The definitive publisher version is available online at 10.1016/j.isprsjprs.2018.04.007

Linear SFM: A Hierarchical Approach to Solving Structure-from-Motion Problems by Decoupling the Linear and Nonlinear Components

Liang Zhao, Shoudong Huang and Gamini Dissanayake

Centre for Autonomous Systems
Faculty of Engineering and Information Technology
University of Technology Sydney, Australia
{*Liang.Zhao,Shoudong.Huang,Gamini.Dissanayake*}*@uts.edu.au*

Abstract

This paper presents a novel hierarchical approach to solving structure-from-motion (SFM) problems. The algorithm begins with small local reconstructions based on nonlinear bundle adjustment (BA). These are then joined in a hierarchical manner using a strategy that requires solving a linear least squares optimization problem followed by a nonlinear transform. The algorithm can handle ordered monocular and stereo image sequences. Two stereo images or three monocular images are adequate for building each initial reconstruction. The bulk of the computation involves solving a linear least squares problem and, therefore, the proposed algorithm avoids three major issues associated with most of the nonlinear optimization algorithms currently used for SFM: the need for a reasonably accurate initial estimate, the need for iterations, and the possibility of being trapped in a local minimum. Also, by summarizing all the original observations into the small local reconstructions with associated information matrices, the proposed linear SFM manages to preserve all the information contained in the observations. The paper also demonstrates that the proposed problem formulation results in a sparse structure that leads to an efficient numerical implementation. The experimental results using publicly available datasets show that the proposed algorithm yields solutions that are very close to those obtained using a global BA starting with an accurate

initial estimate. The C/C++ source code of the proposed algorithm is publicly available at `https://github.com/LiangZhaoPKUImperial/LinearSFM`.

Keywords: Structure-from-motion, local reconstruction, linear least squares, information matrix, monocular, stereo

---

1. Introduction

The structure-from-motion (SFM) problem refers to the process of estimating the three-dimensional structure and the camera trajectory from a set of images [1]. In this paper, the structure is represented by sparse point features and both feature positions and camera poses are estimated using measurements (in the form of feature correspondences) from a sequence of images [2].

Bundle adjustment (BA) can be used to optimize the trajectory and structure by minimizing the re-projection errors [3] once the feature matching is done. In general, BA easily converges to the globally optimal solution for small scale problems. However, BA can be very time-consuming and difficult to converge in large scale problems, unless a good initial estimate is available.

In order to get a good initial estimate and improve the efficiency of the global BA, hierarchical BA is proposed in [4] where BA is performed in a hierarchical way. That is, each level of BA is performed by using the initialization obtained from the results of two smaller BAs in the previous level. This improves the convergence of BA but can still experience problems, especially when a large loop is closed, as will be shown in the experimental results in this paper. Skeletal graphs are proposed in [2], where a small skeletal subset of images is used to reconstruct the skeletal set and then the remaining leaf images are added using pose estimation. This strategy is used in [5][6] for city-scale 3D reconstruction where a conjugate gradient method is applied to solve the linear equations involved in the BA algorithm. To further improve the efficiency of large-scale BA, exact minimum degree ordering and block-based preconditioned conjugate gradient are proposed in [7] and subgraph-preconditioned conjugate gradient is proposed in [8]. However, local optimization methods for the global BA often

require a large number of iterations to converge [5] even if a good initial estimate is available. To improve the convergence of BA and avoid local minima, convex relaxation is proposed in [9] to solve the non-convex optimization problem, although the computational cost is high for large-scale problems.

First building small-scale local reconstructions (submaps) and then combining them to build the global reconstruction is another efficient way to solve large-scale SFM problems [10][11][12][13][14]. For example, in [12], a number of local reconstructions are first independently optimized, then the variables in the local reconstructions not directly used in the merging of reconstructions are factored out to speed up the local reconstructions joining process. In [13], the relative scales between local reconstructions are implicitly included in the state vector of the global map and are optimized through the nonlinear least squares optimization based submap joining process. In [14], reliable triplets are first built and then combined to image sets by hierarchical merging. The focus is the efficiency of the merging of the triplets through the reduction of the number of merged points. In all the above local reconstructions based algorithms, the process of combining the local reconstructions is a nonlinear optimization problem and the initialization is a critical issue that requires further investigation [12].

This paper proposes a new approach to solving SFM problems with ordered image sequences by combining small local reconstructions. In the proposed algorithm, the only part requiring nonlinear optimization is the building of the initial reconstructions by BA. This involves reconstruction from only three monocular images or two stereo images to obtain the camera motion and feature locations. Solution to the complete SFM problem can then be obtained by joining these initial reconstructions by a hierarchical or divide-and-conquer (D&C) [15] process. Each step of joining two local reconstructions in the hierarchical process only requires (i) solving a linear least squares problem and (ii) performing a nonlinear transformation. The proposed algorithm mainly uses linear least squares, therefore, avoids the initialization, iterations and convergence issues involved in most of the nonlinear optimization based algorithms currently used for SFM. Thus it is named Linear SFM.

In the proposed algorithm, the original observations contained in the images are first summarized in the initial reconstructions and the associated information matrices. It, therefore, solves an approximation to the original global BA problem. However, the use of information matrices as the weights in the least squares optimization steps ensures that there is no information loss due to this approximation. Furthermore, the poses and features are always optimized together (the poses and features are correlated through the information matrix). Thus the results obtained using Linear SFM are very close to those using global BA, as clearly seen through the solutions using publicly available datasets. Furthermore, experimental results also demonstrate that Linear SFM has superior capability for closing loops as compared with its nonlinear counterpart.

In a recent work [16], a linear method for global camera pose registration is proposed. The method minimizes an approximate geometric error to enforce the triangular relationship in camera triplets. It has been shown in [16] that the results obtained from different datasets are accurate for point triangulation and can serve as a good initialization for final BA. This linear method is compared with our Linear SFM in this paper and it is shown that our linear approach, although slightly slower, achieves more accurate results.

This paper is based on our preliminary work published as a conference paper [17]. The major improvements of this paper over [17] are: (i) More experimental results using six more datasets (four aerial photogrammetric datasets, KITTI dataset and New College stereo dataset); (ii) Comparison with the linear method proposed in [16]; (iii) Comparison with the hierarchical BA proposed in [4]; (iv) Efficient numerical implementation based on the special sparse structure and the availability of the C/C++ source code at `https://github.com/LiangZhaoPKUImperial/LinearSFM`; (v) An application to stereo vision.

The paper is organized as follows. Section 2 presents the framework of the proposed linear algorithm for monocular SFM. Some details of the linear monocular SFM algorithm are presented in Section 3 and the application to stereo SFM is briefly discussed in Section 4. Section 5 gives the sparse mathematical computation involved in the proposed Linear SFM algorithm for an

efficient implementation. Section 6 presents the results of the proposed linear algorithm using publicly available datasets. Some discussions are elaborated in Section 7. Finally, Section 8 concludes the paper.

2. Framework of Linear Monocular SFM

2.1. Building Initial Reconstructions

The building of a sequence of small initial reconstructions (submaps) is the only nonlinear optimization part involved in the proposed linear monocular SFM algorithm. Thus, we propose to build these submaps as small as possible and call them initial reconstructions. In this paper, each initial reconstruction is built with three images, and there are two common camera poses between two adjacent initial reconstructions [1]. The reason for having two common poses is to allow for the determination of the relative scale between two adjacent local reconstructions (see L_1^1 and L_2^1 in Figure 1).

In order to achieve the best quality of the initial reconstructions, BA is used to build the initial reconstructions. When building the initial reconstructions using BA, we have the flexibility of choosing different coordinate frames and different scale values. More details are given in Section 3.1.

An initial reconstruction can be represented by

$$L_i^1 = (\hat{\mathbf{X}}_i, I_{X_i}). \tag{1}$$

Here an initial reconstruction includes not only the optimal estimate $\hat{\mathbf{X}}_i$ of the state vector from BA, but also the corresponding information matrix I_{X_i} which represents the uncertainty of the initial reconstruction and plays a critical role in our proposed approach.

[1] Here we assume non-zero translation between every two poses; if this is not the case, four or more images are needed to build the initial reconstruction to ensure there are two common poses with non-zero translation between each two adjacent initial reconstructions.

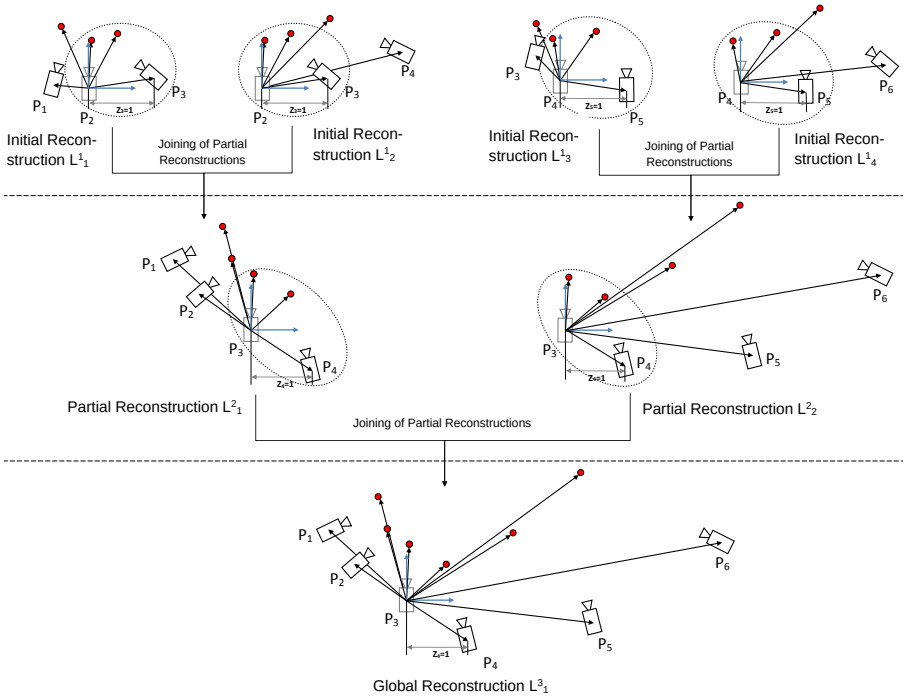


Figure 1: The proposed hierarchical joining of local reconstructions. L_j^i means the jth local reconstruction at level i. Note that the coordinate frames of the local reconstructions at each level are judiciously selected. The poses and features in the circles are the common poses and features between two local reconstructions.

2.2. Solving the SFM in a Hierarchical Manner

After building the initial reconstructions, the global reconstruction can be obtained by hierarchically [18] joining these initial reconstructions.

The process is illustrated in Figure 1. It can be seen that in the proposed hierarchical method only two local reconstructions are joined at each step. As the level increases in the hierarchical process, the size of the two local reconstructions to be joined becomes larger and larger, and the global reconstruction is obtained at the final step of the hierarchical process.

A key point to note here is that the corresponding information matrix is computed in the reconstruction at each level of the hierarchical process and is used to weight the least squares optimization.

2.3. Joining Two Local Reconstructions

As the problem of joining two local reconstructions is similar at each level of the hierarchical process, here we consider one step of joining two initial reconstructions from Figure 1. Suppose the two initial reconstructions to be joined are L_1^1 and L_2^1 as shown in Figure 1. Here L_1^1 is built by using the projections from Images 1, 2 and 3, and L_2^1 is built by using the projections from Images 2, 3 and 4.

Suppose the two local reconstructions L_1^1 and L_2^1 are given by

$$L_1^1 = (\hat{\mathbf{X}}_1,\ I_{X_1}), \quad L_2^1 = (\hat{\mathbf{X}}_2,\ I_{X_2}) \tag{2}$$

and the joint local reconstruction L_1^2 is denoted as

$$L_1^2 = (\hat{\mathbf{X}},\ I_X). \tag{3}$$

The joining of L_1^1 and L_2^1 to build L_1^2 can then be formulated as an optimization problem that

$$\text{argmin}\, \|\hat{\mathbf{X}}_1 - f_1(\mathbf{X})\|^2_{I_{X_1}} + \|\hat{\mathbf{X}}_2 - f_2(\mathbf{X})\|^2_{I_{X_2}} \tag{4}$$

where $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$ are the nonlinear functions relating the state vector $\mathbf{X}$ to the state vectors in the two local reconstructions, and

$$\|\hat{\mathbf{X}}_i - f_i(\mathbf{X})\|^2_{I_{X_i}} = (\hat{\mathbf{X}}_i - f_i(\mathbf{X}))^T I_{X_i} (\hat{\mathbf{X}}_i - f_i(\mathbf{X})), \quad i = 1, 2. \tag{5}$$

When combining the two local reconstructions, L_1^1 and L_2^1 are used as two integrated measurements to build L_1^2 [19]. Since the information matrices I_{X_1} and I_{X_2} represent the uncertainty of the reconstruction estimates $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$, they are used as the weights in the least squares problem (4).

The interesting point is the following: in this paper, we will show that, because two local reconstructions L_1^1 and L_2^1 are in the same coordinate frame with the same scale, the nonlinear part can be decoupled from the original nonlinear optimization problem (4) making the remaining part a linear optimization problem. In other words, the weighted nonlinear least squares problem (4) is equivalent to first solving a weighted linear least squares problem

$$\operatorname{argmin} \|\hat{\mathbf{X}}_1 - A_1\mathbf{Y}\|^2_{I_{X_1}} + \|\hat{\mathbf{X}}_2 - A_2\mathbf{Y}\|^2_{I_{X_2}} \tag{6}$$

and then performing a nonlinear transformation with a closed-form equation

$$\mathbf{X} = \mathbf{Y} \boxplus \Gamma. \tag{7}$$

where A_1 and A_2 are two sparse matrices and Γ is the closed-form nonlinear transformation operation. The details are given in Section 3.2 and Section 3.3.

Remark 1. The information matrices I_{X_1} and I_{X_2} in (4) and (6) play very important roles in the proposed Linear SFM algorithm. (i) All the information involved in the previous local reconstructions is summarized in the local construction estimate and the information matrix, thus, arguably no information is lost in the hierarchical joining process; (ii) The information matrix contains the correlation among all the poses and features in the local reconstruction, thus, all the parameters will be adjusted at once during the optimization.

3. Details of Linear Monocular SFM

This section provides some details of the proposed linear monocular SFM, especially the building and joining of the initial reconstructions.

3.1. Building Initial Reconstructions by BA

The original observations of monocular SFM are the feature projections in the images. Each initial reconstruction is built using the projections from three

images by estimating the camera poses and feature positions. Since only three camera poses are involved, an initial estimate can be easily obtained, such as using the five-point algorithm [20][21]. To ensure the high quality of the initial reconstructions, ParallaxBA [22] is employed to build the initial reconstructions which are then transformed into an XYZ presentation. ParallaxBA has better convergence properties compared to BA using XYZ parametrization [22].

For an initial reconstruction $L_i^1 = (\hat{\mathbf{X}}_i, I_{X_i})$, after obtaining the optimal estimate $\hat{\mathbf{X}}_i$ of the state vector $\mathbf{X}_i$ in the BA, the corresponding information matrix can be obtained by $I_{X_i} = J^T J$ where J is the Jacobian of all the projections in BA, evaluated at the optimal estimate of the state vector.

When performing BA, seven degrees of freedom (DoF), namely six DoF for coordinate frame and one DoF for scale, must be fixed [23]. The rotation and translation of one pose can be fixed as $\mathbf{0}$ to define the coordinate frame, while one more variable needs to be fixed as the scale. Typically, for an exploration trajectory of a camera, the translation in the Z direction is the largest element in the translation vector. In this case the z value of the translation from one pose to another pose can be fixed as 1 to define the scale [2].

When building initial reconstructions by using BA, if the first pose $\mathbf{P}_1$ is fixed as $\mathbf{0}$ to define the coordinate frame, and the z value of the translation from the first pose $\mathbf{P}_1$ to the second pose $\mathbf{P}_2$ is fixed as 1 to define the scale of the reconstructions, we obtain reconstruction $L_i^1 = (\hat{\mathbf{X}}_i, I_{X_i})$ in Figure 2. If the second pose $\mathbf{P}_2$ is fixed as $\mathbf{0}$ to define the coordinate frame, and the z value of the translation from the second pose $\mathbf{P}_2$ to the third pose $\mathbf{P}_3$ is fixed as 1 to define the scale of the reconstructions, we get reconstruction $L_i^{1\prime} = (\hat{\mathbf{X}}_i', I'_{X_i})$ in Figure 2.

Note that the two reconstructions in Figure 2, L_i^1 and $L_i^{1\prime}$, are equivalent because they are both optimal solutions obtained using the same amount of

[2] In this paper, we use the case of fixing the z value as an example. There will be other cases where x or y is the largest value (e.g. in aerial photogrammetric datasets) but the methods are similar.

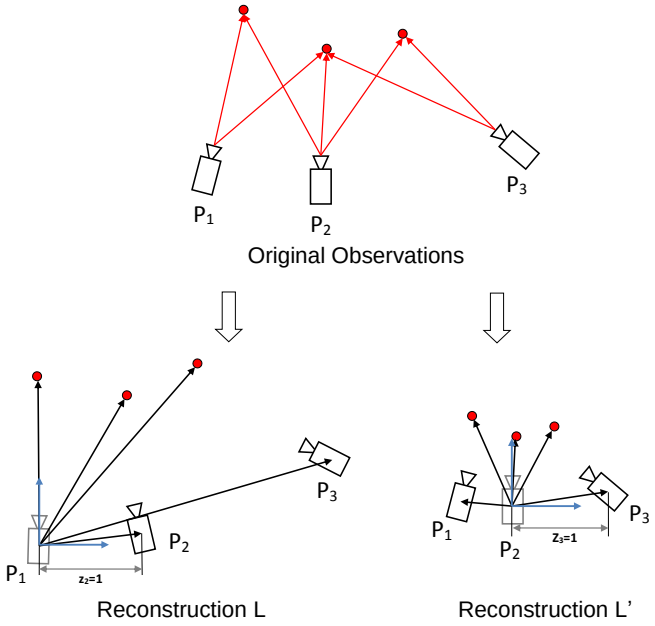


Figure 2: Building different local reconstructions from the same observations by using different coordinate frames and different scales. The two local reconstructions are equivalent.

information. In the proposed Linear SFM, the coordinate frames and scales are judiciously selected when building the local reconstructions, as shown in Figure 1.

3.2. Joining Two Local Reconstructions Step I: Linear Least Squares

Note that the two initial reconstructions are in the same coordinate frame with the same scale, thus, the joining of the two local reconstructions can be separated into two steps: (i) solving linear least squares, and (ii) performing a nonlinear transformation.

Suppose there are two local reconstructions L_1^1 and L_2^1 given in (2), where the state vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ of local reconstructions L_1^1 and L_2^1 are defined as (for simplicity, some transposes of vectors are omitted in this paper)

$$\begin{aligned} \mathbf{X}_1 &= [{}^2\mathbf{r}_1,\ {}^2\mathbf{t}_1,\ {}^2\mathbf{F}_1,\ {}^2\mathbf{r}_3,\ {}^2x_3,\ {}^2y_3,\ {}^2\mathbf{F}_C] \\ \mathbf{X}_2 &= [{}^2\mathbf{r}_3,\ {}^2x_3,\ {}^2y_3,\ {}^2\mathbf{F}_C,\ {}^2\mathbf{r}_4,\ {}^2\mathbf{t}_4,\ {}^2\mathbf{F}_2] \end{aligned} \tag{8}$$

and I_{X_1} and I_{X_2} are the associated information matrices. [3]

In the state vector $\mathbf{X}_1$ in (8), pose $\mathbf{P}_1$ in the coordinate frame of $\mathbf{P}_2$ is presented by

$$^2\mathbf{P}_1 = [^2\mathbf{r}_1,\ ^2\mathbf{t}_1] \tag{9}$$

where $^2\mathbf{r}_1 = [^2\alpha_1\ ^2\beta_1\ ^2\gamma_1]^T$ is the vector containing the three Euler angles, and $^2\mathbf{t}_1 = [^2x_1,\ ^2y_1,\ ^2z_1]^T$ is the translation; $^2\mathbf{F}_1$ and $^2\mathbf{F}_C$ represent all the feature XYZ positions in L_1^1.

Here $\mathbf{P}_2 = \mathbf{0}$ is fixed as the coordinate frame and $^2z_3 = 1$ is fixed as the scale for both L_1^1 and L_2^1, thus, they are not in the state vectors $\mathbf{X}_1$ and $\mathbf{X}_2$. Obviously, L_1^1 and L_2^1 are in the same coordinate system.

In the state vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ in (8), $^2\mathbf{F}_1$,$^2\mathbf{F}_C$,$^2\mathbf{F}_2$ are used to represent the features, where $^2\mathbf{F}_C$ represents the common features that appear in both of the two reconstructions, while $^2\mathbf{F}_1$ ($^2\mathbf{F}_2$) represents the features that only appear in L_1^1 (L_2^1).

We denote the intermediate state vector $\mathbf{Y}$ in the joining of two reconstructions in (6) as

$$\mathbf{Y} = [^2\mathbf{r}_1,\ ^2\mathbf{t}_1,\ ^2\mathbf{F}_1,\ ^2\mathbf{r}_3,\ ^2x_3,\ ^2y_3,\ ^2\mathbf{F}_C,\ ^2\mathbf{r}_4,\ ^2\mathbf{t}_4,\ ^2\mathbf{F}_2]. \tag{10}$$

Because $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{Y}$ are all in the same coordinate system, the observation functions $f_1(\cdot)$ and $f_2(\cdot)$ in (4) in Section 2.3 become linear if $\mathbf{Y}$ is used as the state vector. Thus, the joining of two reconstructions becomes a linear least squares problem as in (6) where the coefficient matrices A_1 and A_2 are formed by identity and zero matrices as follows

$$A_1 = \left[\mathrm{E}_{n_1} \mid 0_{n_1 \times m_2}\right], \quad A_2 = \left[0_{n_2 \times m_1} \mid \mathrm{E}_{n_2}\right]. \tag{11}$$

Here n_1 and n_2 are the dimensions of the state vectors $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively. $m_1 = 3k_1 + 6$ and $m_2 = 3k_2 + 6$, where k_1 and k_2 are the number of features

[3] Here and in the following, a number i at the upper left corner of a variable means the coordinate frame is $\mathbf{P}_i$.

in ${}^2\mathbf{F}_1$ and ${}^2\mathbf{F}_2$, respectively. Thus m_1 (m_2) is the dimension of the poses and features in the state vectors $\mathbf{X}_1$ ($\mathbf{X}_2$) which only appear in L_1^1 (L_2^1).

The optimal solution $\hat{\mathbf{Y}}$ of linear least squares problem (6) can be calculated by solving the sparse linear equation

$$(A^T I_Z A)\hat{\mathbf{Y}} = A^T I_Z \mathbf{Z}. \tag{12}$$

where $\mathbf{Z} = [\hat{\mathbf{X}}_1^T,\ \hat{\mathbf{X}}_2^T]^T$, $I_Z = \mathrm{diag}(I_{X_1}, I_{X_2})$, and $A = [A_1^T,\ A_2^T]^T$.

The corresponding sparse information matrix of $\hat{\mathbf{Y}}$ can be computed as

$$I_Y = A^T I_Z A \tag{13}$$

It is obvious that the linear least squares formulation (6) can be extended to the joining of two larger local reconstructions (for example, L_1^2 and L_2^2 in Figure 1) which are in the same coordinate system.

3.3. Joining Two Local Reconstructions Step II: Nonlinear Transformation

The joint local reconstruction $L_1^2 = (\hat{\mathbf{X}},\ I_X)$ in (3) can then be obtained from the intermediate estimate and information matrix $(\hat{\mathbf{Y}},\ I_Y)$ by a nonlinear transformation with closed-form solution.

Suppose the state vector of the aimed reconstruction L_1^2 is defined as

$$\mathbf{X} = [{}^3\mathbf{r}_1,\ {}^3\mathbf{t}_1,\ {}^3\mathbf{r}_2,\ {}^3\mathbf{t}_2,\ {}^3\mathbf{r}_4,\ {}^3x_4,\ {}^3y_4,\ {}^3\mathbf{F}]. \tag{14}$$

Here $\mathbf{P}_3 = \mathbf{0}$ is fixed as the coordinate frame and ${}^3z_4 = 1$ is fixed as the scale of L_1^2, thus, they are not in the state vector. In the state vector $\mathbf{X}$, for simplification, $\mathbf{F}$ is used instead of $\mathbf{F}_1$, $\mathbf{F}_2$ and $\mathbf{F}_C$ in (10) to represent all the features in the joint reconstruction.

Then, the nonlinear transformation from $\mathbf{Y}$ to $\mathbf{X}$ in (7) can be given by

$$\mathbf{X} = g(\mathbf{Y}) = \mathbf{Y} \boxplus \Gamma$$
$$\Rightarrow \begin{cases} {}^3\mathbf{r}_1 = r^{-1}({}^2R_1 {}^2R_3^T) \\ {}^3\mathbf{t}_1 = {}^2R_3\,({}^2\mathbf{t}_1 - [{}^2x_3, {}^2y_3, 1]^T)/z_s \\ {}^3\mathbf{r}_2 = r^{-1}({}^2R_3^T) \\ {}^3\mathbf{t}_2 = -{}^2R_3\,[{}^2x_3, {}^2y_3, 1]^T/z_s \\ {}^3\mathbf{r}_4 = r^{-1}({}^2R_4 {}^2R_3^T) \\ {}^3x_4 = x_s/z_s \\ {}^3y_4 = y_s/z_s \\ {}^3\mathbf{F} = {}^2R_3\,({}^2\mathbf{F} - [{}^2x_3, {}^2y_3, 1]^T)/z_s \end{cases} \tag{15}$$

where $r(\cdot)$ and $r^{-1}(\cdot)$ are the angle-to-matrix and matrix-to-angle functions, ${}^2R_1 = r({}^2\mathbf{r}_1)$, ${}^2R_3 = r({}^2\mathbf{r}_3)$ and ${}^2R_4 = r({}^2\mathbf{r}_4)$ are the rotation matrices of pose ${}^2\mathbf{P}_1$, ${}^2\mathbf{P}_3$ and pose ${}^2\mathbf{P}_4$ in the state vector $\mathbf{Y}$, respectively.

In (15), the scale factor z_s as well as x_s and y_s can be computed as

$$[x_s,\ y_s,\ z_s]^T = {}^2R_3\,({}^2\mathbf{t}_4 - [{}^2x_3, {}^2y_3, 1]^T). \tag{16}$$

From the intermediate estimates $(\hat{\mathbf{Y}},\ I_Y)$, the estimate of the state vector $\mathbf{X}$ in the joint reconstruction L_1^2 can be obtained by

$$\hat{\mathbf{X}} = g(\hat{\mathbf{Y}}) = \hat{\mathbf{Y}} \boxplus \Gamma. \tag{17}$$

The corresponding information matrix I_X is given by

$$I_X = \nabla^T I_Y \nabla \tag{18}$$

where ∇ is the Jacobian of $\mathbf{Y}$ with respect to $\mathbf{X}$, evaluated at $\hat{\mathbf{X}}$

$$\nabla = \frac{\partial g^{-1}(\mathbf{X})}{\partial \mathbf{X}}|_{\hat{\mathbf{X}}} \tag{19}$$

with

$$\mathbf{Y} = g^{-1}(\mathbf{X}) = \mathbf{X} \boxminus \Gamma \tag{20}$$

the inverse function of $g(\cdot)$ in (15).

The reason why $g(\cdot)$ has closed form is that the relation between the intermediate state vector $\mathbf{Y}$ and the aimed state vector $\mathbf{X}$ is the coordinate and scale transformation function, transforming $\mathbf{Y}$ in the coordinate frame of $\mathbf{P}_2$ with the scale ${}^2z_3 = 1$, into $\mathbf{X}$ in the coordinate frame of $\mathbf{P}_3$ with the scale ${}^3z_4 = 1$. Thus, both the state estimate and the corresponding information matrix can be transformed easily using closed-form formulas.

3.4. Joining a Sequence of Initial Reconstructions in a Hierarchical Manner

Based on the joining of two local reconstructions, joining a sequence of initial reconstructions to obtain the global reconstruction can be done in a hierarchical manner.

As can be seen from Figure 1, both the two local reconstructions L_1^1 and L_2^1 are in the coordinate frame defined by $\mathbf{P}_2$, with the scale defined by ${}^2z_3 = 1$. They are joined to build local reconstructions L_1^2 which is in the coordinate frame defined by $\mathbf{P}_3$, with the scale defined by ${}^3z_4 = 1$. Similarly, L_3^1 and L_4^1 are in the coordinate frame of $\mathbf{P}_4$, with the scale ${}^4z_5 = 1$. The joint reconstructions L_2^2 is in the coordinate frame of $\mathbf{P}_3$, with the scale ${}^3z_4 = 1$. Then the two local reconstructions L_1^2 and L_2^2 at the second level in the hierarchical process are in the same coordinate frame with the same scale, thus, joining L_1^2 and L_2^2 to build the reconstruction L_1^3 can be done the same way as joining L_1^1 and L_2^1 to build L_1^2.

Since the two local reconstructions to be joined are always in the same coordinates with the same scale, the nonlinear joining can be formulated as a linear least squares problem and a nonlinear transformation as described in Sections 3.2 and 3.3. Thus, the whole SFM problem can be solved by joining a number of initial reconstructions (e.g. L_1^1, L_2^1, L_3^1, and L_4^1 in Figure 1) to build the global reconstruction by a hierarchical method, with only linear least squares and nonlinear coordinate and scale transformations.

As stated above, the key ideas of the proposed Linear SFM algorithm are that: (i) A local reconstruction can be built in different coordinate frames with

different scales by fixing different camera poses within the local reconstruction; (ii) By judiciously selecting the coordinate frames and scales of the local reconstructions, at each step in the hierarchical process, the two local reconstructions to be joined can be in the same coordinate frame with the same scale; and (iii) since the two local reconstructions are in the same coordinate frame with the same scale, the nonlinear joining of these two local reconstructions can be decoupled by a linear least squares optimization and a nonlinear coordinates and scale transformation. Algorithm 1 describes the hierarchical process of the proposed Linear SFM algorithm.

Algorithm 1 The Hierarchical Process in Linear SFM

Input: sequence of initial reconstructions $\{L_j^1\}$

Output: the global reconstruction L_1^m

Traverse each level in the hierarchical process
for $i = 1; i <= m; i++$ **do**
  Traverse each pair of local reconstructions at each level
  for $j = 1; j <= n_i; j = j + 2$ **do**
    Joining L_j^i and L_{j+1}^i to get $L_{(j+1)/2}^{i+1}$ (NONLINEAR)
    (i) Solve least squares (6) to get $(\hat{\mathbf{Y}},\ I_Y)$ (LINEAR)
    (ii) Apply transformation (15) and (18) to get $L_{(j+1)/2}^{i+1}$ (NONLINEAR)
  end for
end for

4. An Application to Stereo SFM

Since the scale is known in stereo vision, each initial reconstruction can be built using images from only two camera poses and one camera pose serves as the coordinate frame. There is one common pose between every two adjacent initial reconstructions such that the reconstructions can be linked.

Different from (8) for monocular SFM in Section 3.3, the state vector of

initial reconstruction L_1^1 is given by

$$\mathbf{X}_1 = [{}^1\mathbf{r}_2,\ {}^1\mathbf{t}_2,\ {}^1\mathbf{F}]. \tag{21}$$

Here the pose is in 6 DoF because no element is fixed as scale.

The joining of two local reconstructions can be implemented in a way similar to the method described in Sections 3.2 and 3.3, where in the nonlinear transformation, the scale factor in (15) can be ignored because the scale is observable and all the local reconstructions have the same scale. Then the process of joining a sequence of initial reconstructions to build the global reconstruction in the hierarchical manner is the same as that of Linear SFM for monocular vision.

5. Exploiting the Sparsity in Linear SFM

The computational cost is one of the most important issues in SFM, since the problems often involve thousands of poses and hundreds of thousands of features. There are several efficient BA implementations, e.g. SBA [3], sSBA [24], g2o [25], multicore BA (PBA) [26] and ParallaxBA [13][27], based on the special structure of the sparse matrices in the SFM problems. In this section, we will show that instead of using general sparse matrix libraries, the proposed Linear SFM algorithm can also be implemented in an efficient way by exploiting the sparse structure, which is different from that in BA.

In the process of the proposed Linear SFM algorithm, the most computationally costly parts are (i) transforming the information matrix (18), (ii) constructing the linear equation (12), and (iii) solving the linear equation (12). In the following, the implementation for these three parts are described in details.

5.1. Transformation of Information Matrix

This subsection describes how to implement the transformation of the information matrix in (18). For simplification, suppose I (I_Y in (18)) is the original information matrix, I' (I_X in (18)) is the transformed information matrix, and ∇ is the Jacobian of the transformation. An example structure of the three

matrices could be

$$I = \begin{bmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & \bullet & \\ & \bullet & \bullet & \bullet & \\ & \bullet & \bullet & \bullet & \bullet \\ & & & \bullet & \bullet \end{bmatrix}, \nabla = \begin{bmatrix} \bullet & \bullet & & & \\ & \bullet & & & \\ & \bullet & \bullet & & \\ & \bullet & & \bullet & \\ & \bullet & & & \bullet \end{bmatrix}, I' = \begin{bmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \\ & \bullet & \bullet & \bullet & \bullet \\ & \bullet & & \bullet & \bullet \end{bmatrix} \tag{22}$$

where $\bullet$ represents a nonzero block in the sparse matrices.

In the proposed Linear SFM algorithm, the Jacobian of the transformation ∇ is a block diagonal sparse matrix, except for one full column of blocks. The latter corresponds for example in the stereo case to the pose in the state vector $\mathbf{X}$ in (14) which defined the coordinate frame of the state vector $\mathbf{Y}$ in (10).

The Jacobian can be decoupled by $\nabla = \nabla_1 + \nabla_2$, where ∇_1 is block diagonal and ∇_2 is the block column of the Jacobian ∇, without the block on the diagonal. For example, for ∇ in (22), we have

$$\nabla_1 = \begin{bmatrix} \bullet & & & & \\ & \bullet & & & \\ & & \bullet & & \\ & & & \bullet & \\ & & & & \bullet \end{bmatrix}, \nabla_2 = \begin{bmatrix} \bullet & \\ & \\ \bullet & \\ \bullet & \\ \bullet & \end{bmatrix}. \tag{23}$$

Suppose there are n blocks in each row and column in I and I', i and j are the row and column ID for a nonzero block in the sparse matrices, k is the column ID for the nonzero block column in ∇, $\nabla_1(i)$ and $\nabla_2(j)$ are the i^{th} and j^{th} blocks in ∇_1 and ∇_2, respectively. Thus, the (k,k) block in the transformed information matrix I' can be computed as

$$\begin{aligned} I'(k,k) = {} & \nabla_1^T(k)\, I(k,k)\, \nabla_1(k) + \sum^{i,j} \nabla_2^T(i)\, I(i,j)\, \nabla_2(j) \\ & + \sum^{i} \nabla_2^T(i)\, I(i,k)\, \nabla_1(k) + \sum^{j} \nabla_1^T(k)\, I(k,j)\, \nabla_2(j) \end{aligned} \tag{24}$$

where

$$\sum^{i} \nabla_2^T(i)\, I(i,k)\, \nabla_1(k) = \left(\sum^{j} \nabla_1^T(k)\, I(k,j)\, \nabla_2(j) \right)^T \tag{25}$$

since both information matrices I and I' are symmetric.

For the blocks on the k^{th} column in I',

$$I'(i,k) = \sum^{j} \nabla_1^T(i)\, I(i,j)\, \nabla_2(j), \qquad i \neq k \tag{26}$$

and for the blocks on the k^{th} row in I' holds,

$$I'(k,j) = \sum^{i} \nabla_2^T(i)\, I(i,j)\, \nabla_1(j), \qquad j \neq k \tag{27}$$

where if $i = j$, $I'(i,k) = (I'(k,j))^T$.

For any other blocks in I' neither on k^{th} column nor on k^{th} row, we obtain

$$I'(i,j) = \nabla_1^T(i)\, I(i,j)\, \nabla_1(j), \qquad i,j \neq k. \tag{28}$$

5.2. Constructing the Linear Equation

In this subsection, we will show the computation of the coefficient matrix $A^T I_Z A$ and the column vector $A^T I_Z \mathbf{Z}$ in the linear equation (12). For simplification, we use $(\mathbf{X}_1,\ I_1)$ and $(\mathbf{X}_2,\ I_2)$ (instead of $(\hat{\mathbf{X}}_1,\ I_{X_1})$ and $(\hat{\mathbf{X}}_2,\ I_{X_2})$) to present the two reconstructions to be joined, and $(\mathbf{X},\ I)$ to represent the joint reconstruction.

The coefficient matrix $A^T I_Z A$ (or the information matrix I of the joint reconstruction) is just the combination of the two information matrices I_1 and I_2 of the two local reconstructions, considering the common blocks corresponding to the common pose and features.

Suppose (i_1, j_1), (i_2, j_2) and (i, j) are the row and column IDs for a nonzero block in the information matrix I_1, I_2 and I, respectively. $u_1(i)$ and $u_2(i)$ are the corresponding ID of i in I_1 and I_2, which correspond to the same pose or feature between the local reconstruction and joint reconstruction. Then each nonzero block in $I = A^T I_Z A$ can be computed as

$$I(i,j) = I_1(u_1(i), u_1(j)) + I_2(u_2(i), u_2(j)) \tag{29}$$

where $I_1(u_1(i), u_1(j)) = 0$ if either $u_1(i)$ or $u_1(j)$ does not exist.

The column vector $\mathbf{b} = A^T I_Z \mathbf{Z}$ can be computed by using the information matrices and the estimates of the state vectors of the two local reconstructions $(\mathbf{X}_1,\ I_1)$ and $(\mathbf{X}_2,\ I_2)$ as

$$\mathbf{b}(i) = \sum^{j_1} I_1(u_1(i), j_1)\, \mathbf{X}_1(j_1) + \sum^{j_2} I_2(u_2(i), j_2)\, \mathbf{X}_2(j_2) \tag{30}$$

where $\mathbf{b}(i)$ represents the block in vector $\mathbf{b}$ corresponding to the i^{th} row blocks in I.

5.3. Solving the Linear Equation

The linear equation (12) in Section 3.2 can be solved in a similar way as the one in each iteration of SBA [3] or ParallaxBA [22].

Suppose the linear equation (12) is rewritten as

$$\begin{bmatrix} U & W \\ W^T & V \end{bmatrix} \begin{bmatrix} \mathbf{X}_P \\ \mathbf{X}_F \end{bmatrix} = \begin{bmatrix} \mathbf{b}_P \\ \mathbf{b}_F \end{bmatrix} \tag{31}$$

where $\mathbf{X}_P$ and $\mathbf{X}_F$ are the pose and feature parts of the state vector $\mathbf{X}$. Since in the SFM problem, the number of features is much larger than the number of poses, the whole linear system (31) can be solved by first solving the reduced camera system using Schur complement

$$(U - WV^{-1}W^T)\mathbf{X}_P = \mathbf{b}_P - WV^{-1}\mathbf{b}_F \tag{32}$$

and then performing back-substitution to get the features

$$V\mathbf{X}_F = \mathbf{b}_F - W^T\mathbf{X}_P. \tag{33}$$

As matrix V is block diagonal, its inverse can be easily computed. Thus, an efficient implementation can be achieved similar to ParallaxBA [22][25]. Note that here the camera poses and features are directly solved in one step without any iterations, which is different from nonlinear optimization based BA.

We finally note that, similar to BA, all the information matrices, e.g. I, I', I_1 and I_2 are symmetric matrices. Thus, only the upper triangle of the information matrix is stored and calculated in the proposed algorithm and implementation.

In summary, as addressed in this section, the computational complexity of the above three parts in the proposed Linear SFM algorithm is $\Theta(m)$, where m is the number of the nonzero blocks in the upper/lower triangle of the information matrix I.

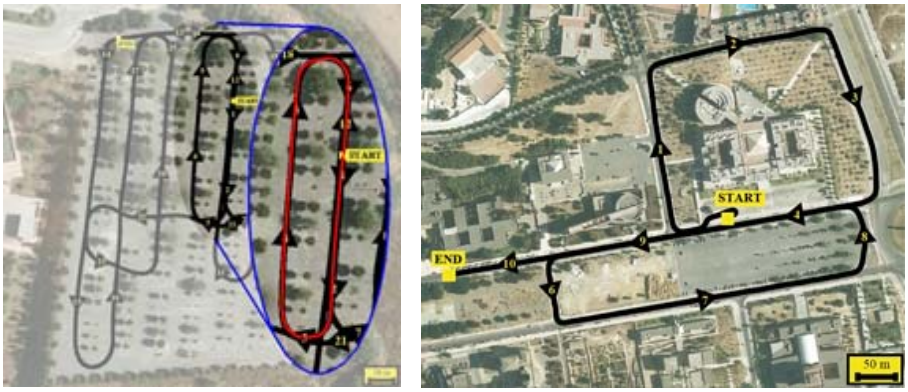
6. Experimental Results

In this section, the publicly available "Malaga" [28], "KITTI" [29], and "Aerial Photogrammetry" (AP) [22][27][30][31] datasets are used for monocular Linear SFM, and the "New College" dataset is employed for stereo Linear SFM to demonstrate that the results by the proposed approach are very close to the ground truth and/or the results obtained by global BA with good initialization. Additionally, a comparison to hierarchical BA [32] has been conducted for monocular SFM using the monocular datasets. It is shown that the nonlinear optimization based BA algorithm in hierarchical BA sometimes fails to converge to the correct solution when a large loop is closed, while the proposed Linear SFM approach achieves good results. Furthermore, a comparison to the linear method proposed in [16] (LinearCamReg) is also performed using five AP datasets. It is shown that our approach produces more accurate results.

6.1. Malaga Datasets

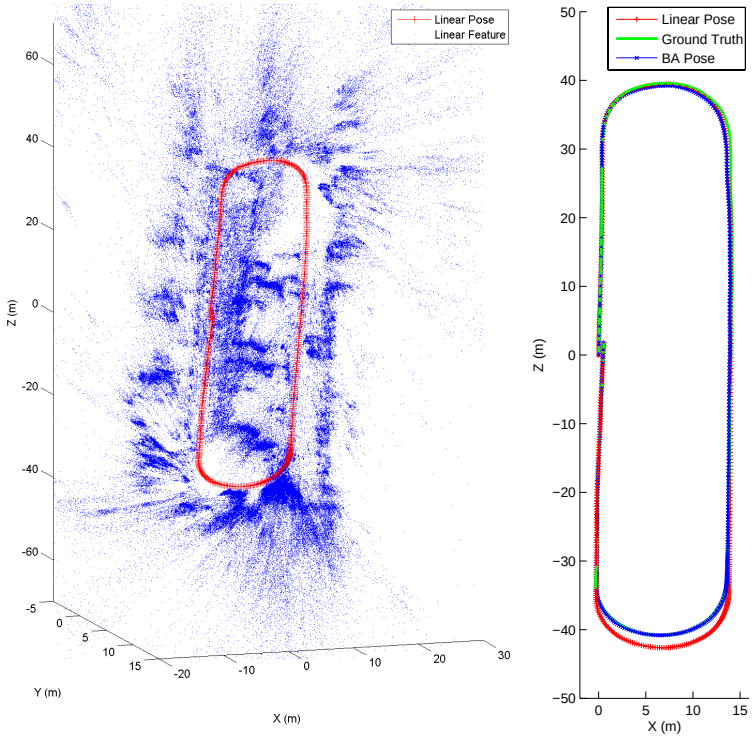
In the Malaga 2009 Robotic Dataset Collection [28], the image resolution is 1024×768 and the camera calibration parameters are provided in the dataset. Images captured by the right camera are used. SIFT [33] and RANSAC [34] are used for feature detection, matching, and outlier removal, including the loop closure detection. In this paper, we use L^2-SIFT [35] which is an efficient implementation to data association based on SIFTGPU, especially for large images and large datasets. The number of features and projections after SIFT matching and RANSAC outlier removal can be found in Table 1.

As described in Section 2.1, three images are used to build each initial reconstruction by ParallaxBA and then transformed into an XYZ presentation.



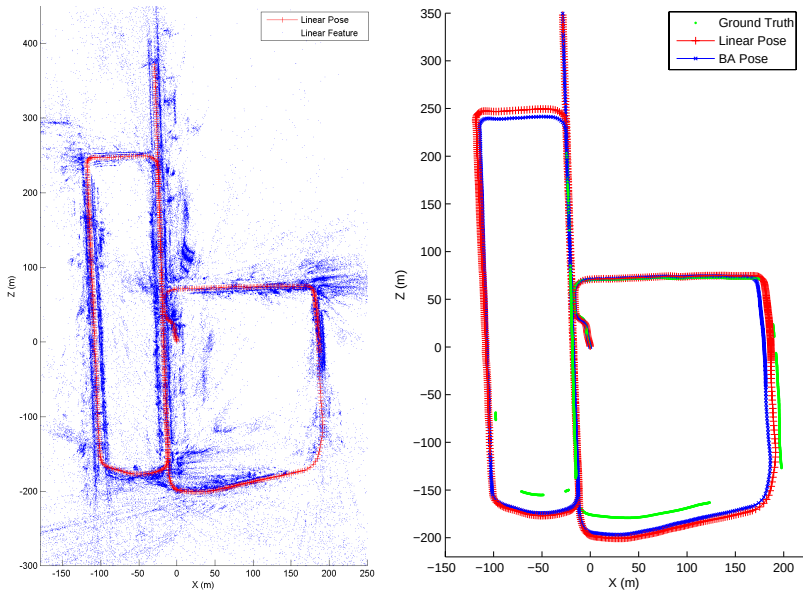
(a) PARKING-6L Dataset (b) CAMPUS-2L Dataset

Figure 3: The trajectories of Malaga monocular datasets.



(a) Pose Feature (b) Comparison

Figure 4: Linear SFM result of PARKING-6L monocular dataset (508 poses, 190,711 features and 567,836 projections).



(a) Pose Feature (b) Comparison

Figure 5: Linear SFM result of CAMPUS-2L monocular dataset (1,020 poses, 198,563 features and 575,644 projections).

The initial estimates are computed by two-view geometry and ParallaxBA takes three to five iterations to converge with the mean square of the final re-projection errors (MSE) around 0.1 pixel^2.

The PARKING-6L dataset contains one sequence of images collected from a 250m closed loop trajectory (Figure 3(a)) with 508 images. The result of Linear SFM is shown in Figure 4(a). The estimated poses are compared with the result of global BA as well as the ground truth in Figure 4(b).

The CAMPUS-2L dataset (Figure 3(b)) has a 2.2km long trajectory with two loops. 1,020 keyframes are selected from 5,103 images by simply selecting one in every five images. The result of the proposed Linear SFM approach, the global BA result and the ground truth are shown in Figure 5(a) and Figure 5(b), respectively.

One step of the hierarchical process is shown in Figure 6. Local reconstruction L_1^9 (the 1^{st} local reconstruction at the 9^{th} level of the hierarchical process) is

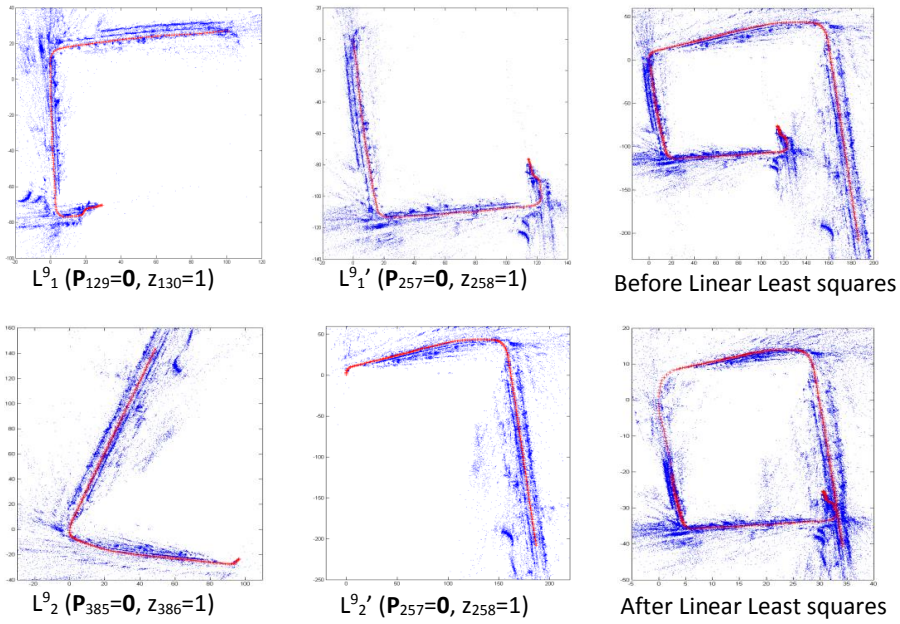


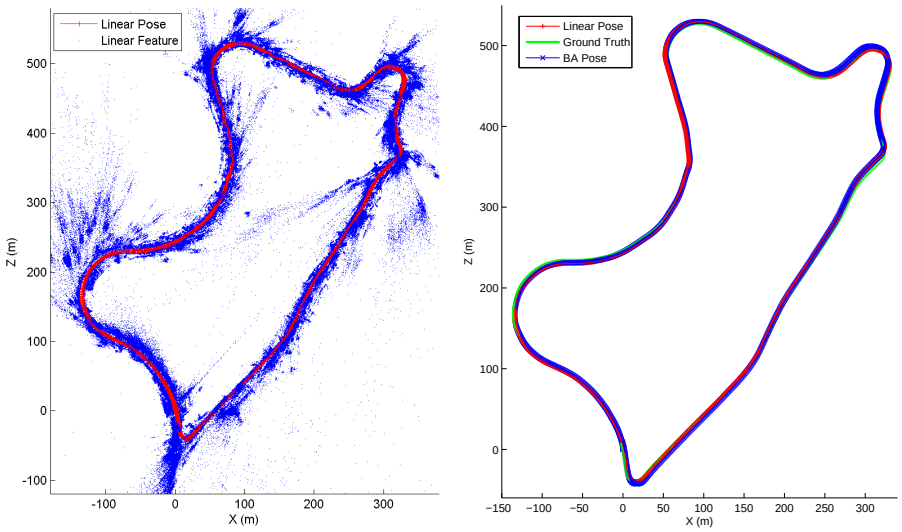
Figure 6: Intermediate step of CAMPUS-2L monocular dataset.

built in the coordinate frame of $\mathbf{P}_{129}$, and local reconstruction L_2^9 is built in the coordinate frame of $\mathbf{P}_{385}$. First, the two local reconstructions are transformed into $L_1^{9'}$ and $L_2^{9'}$ which have the same coordinate frame of $\mathbf{P}_{257}$. As shown in Figure 6, before linear least squares optimization, the two local reconstructions show a large drift; after the linear least squares optimization, a good quality local reconstruction L_1^{10} is built (the 1^{st} local reconstruction at the 10^{th} level of the hierarchical process).

6.2. KITTI Dataset

KITTI 09 in the KITTI Vision Benchmark Suite dataset [29] is chosen with the trajectory of about 1,700m consisting of a sequence of 1591 keyframes. Similar to the Malaga Dataset, high-precision ground truth is available from the GPS/IMU system.

Only images from the left camera are used. There are 1,591 poses, 520,533 features and 1,427,645 projections remaining after L^2-SIFT. 1,589 initial recon-



(a) Pose Feature (b) Comparison

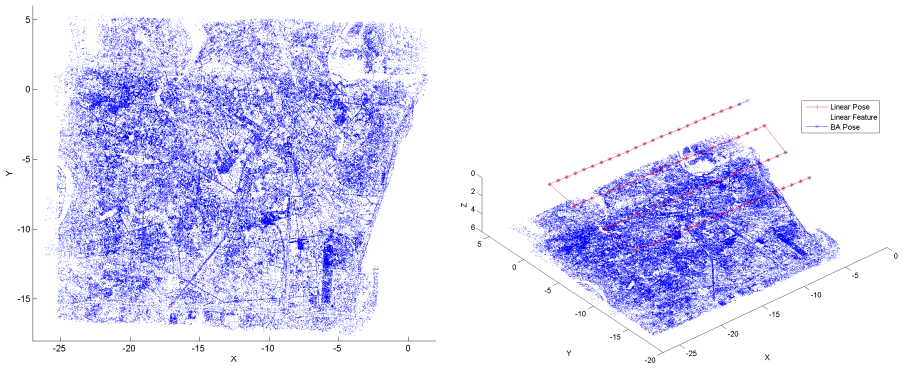
Figure 7: Linear SFM result of KITTI 09 monocular dataset (1,591 poses, 520,533 features and 1,427,645 projections).

structions, each with three images, are built using ParallaxBA. The result of the proposed Linear SFM algorithm is shown in Figure 7(a). As comparison, the ground truth and the result of the global BA are shown in Figure 7(b).

6.3. Aerial Photogrammetric Datasets

Six aerial photogrammetric datasets are also used. For these six datasets, the cameras are mounted on the aerial platforms to map the ground surface.

The Village Dataset contains 90 images taken by digital mapping camera (DMC) in a snake track with image resolution 7680×13824 pixels, mapping an area of about 3.4km $\times$ 2km. The College Dataset contains 468 images with resolution 5616×3744 captured by a Canon camera, mapping a university campus with a size of about 3.0km $\times$ 2.9km. In the Dunhuan and Jinan Datasets, the same Canon cameras are used to capture 63 images with 3 tracks and 76 images with 2 tracks, respectively. Vaihingen and Toronto datasets are from the ISPRS Test Project on Urban Classification and 3D Building Reconstruction [30][31].



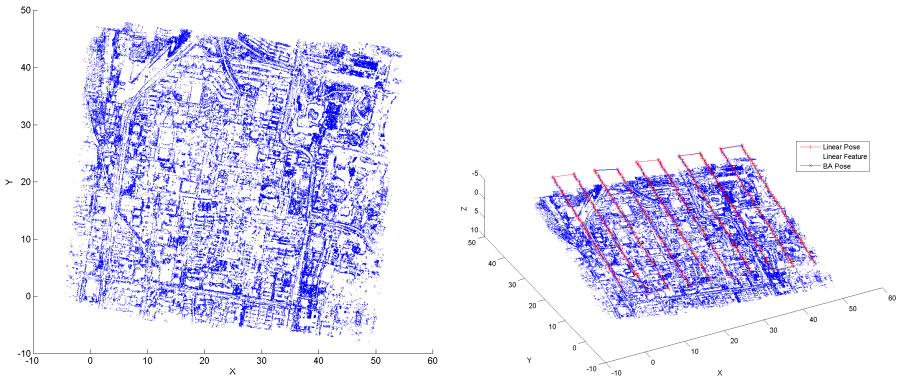
(a) Top view (b) 3D view

Figure 8: Linear SFM result of Village dataset (90 poses, 273,131 3D features and 779,268 projections).

In the Vaihingen dataset, there are 20 images with a resolution of 7680×13824 pixels in three tracks captured by a DMC camera. For the Toronto dataset, there are 13 images with image resolution 11500×7500 pixels in three tracks captured by a UCD camera. L^2-SIFT feature extraction and outlier removal is used for all six datasets and the data association results are summarized in Table 1. The mapping results as well as the camera poses by Linear SFM are shown in Figures 8-13. As comparison, the camera pose estimates by global BA are given in Figures 8(b)-13(b) (the two results of camera poses overlap each other).

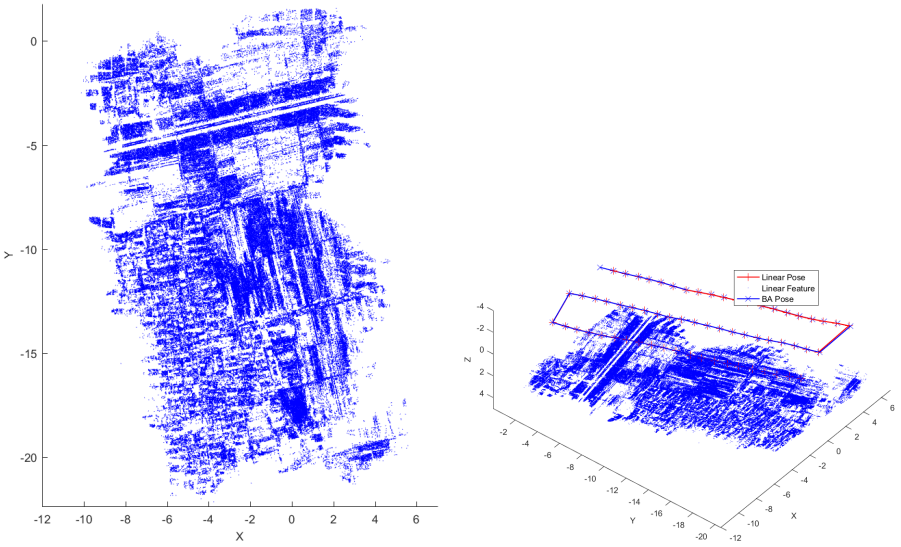
6.4. Comparison to Hierarchical BA

As comparison, the hierarchical BA [4][32] is also applied to the above nine monocular datasets. For the hierarchical BA, BA is performed in the same hierarchical way as in Linear SFM. The hierarchical BA is also started from every three images. After that, each BA is performed by using the observations from the previous two BAs, and initialized based on the results from the previous two BAs. In the initialization, the two estimates of the state vectors in the previous BAs are linked, and the relative scale is derived from the two common poses. The Levenberg-Marquardt algorithm is used to solve the nonlinear optimization



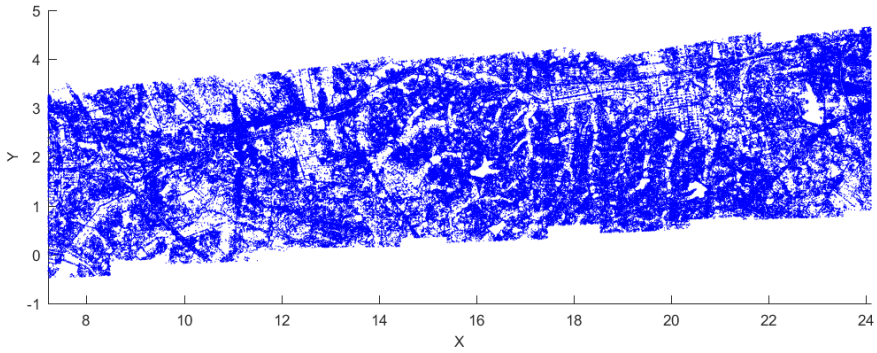
(a) Top view (b) 3D view

Figure 9: Linear SFM result of College dataset (468 poses, 444,596 3D features and 1,368,258 projections).

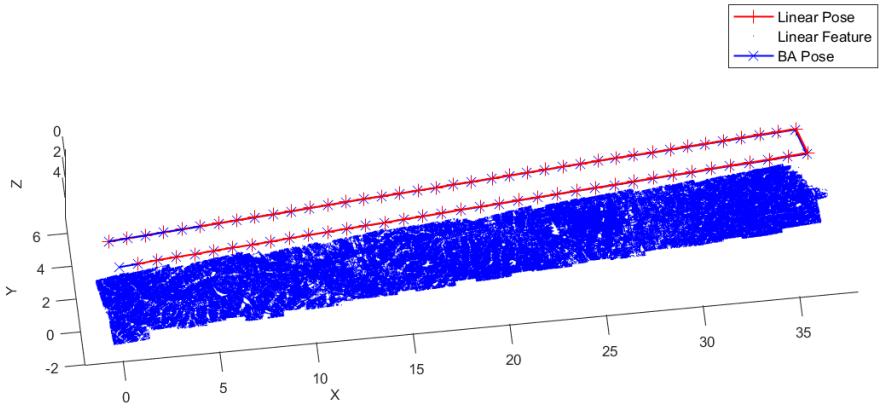


(a) Top view (b) 3D view

Figure 10: Linear SFM result of Dunhuan dataset (63 poses, 250,782 3D features and 597,289 projections).

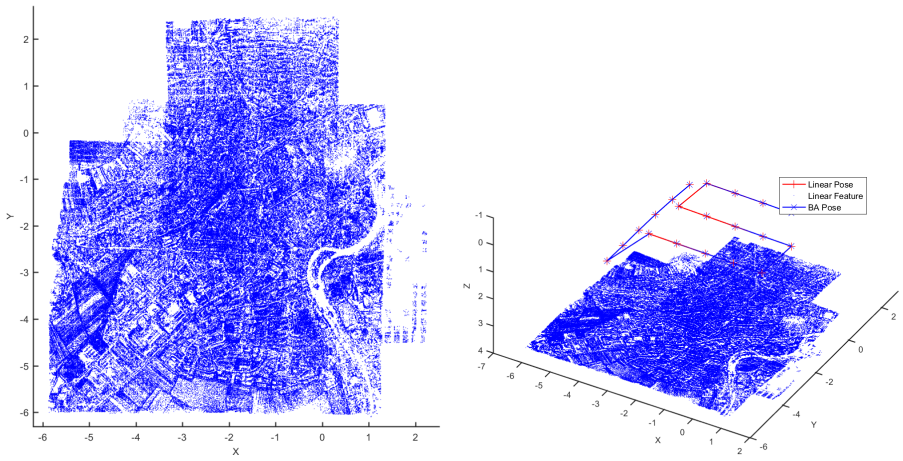


(a) Top view



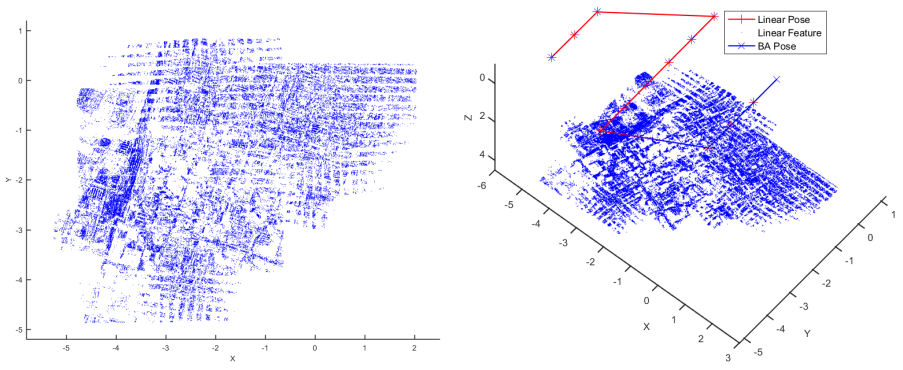
(b) 3D view

Figure 11: Linear SFM result of Jinan dataset (76 poses, 1,228,959 3D features and 2,864,740 projections).



(a) Top view (b) 3D view

Figure 12: Linear SFM result of Vaihingen dataset (20 poses, 554,169 3D features and 1,201,982 projections).



(a) Top view (b) 3D view

Figure 13: Linear SFM result of Toronto dataset (13 poses, 113,685 3D features and 239,279 projections).

in each BA and the maximum iteration number is set as 100.

Table 1: Computational Costs* of Linear SFM Algorithm (in seconds)

Dataset	Poses	Features	Projections	time
PARKING-6L	508	190711	567836	29.736
CAMPUS-2L	1020	198563	575644	47.688
KITTI 09	1591	520533	1427645	122.37
AP Village	90	273131	779268	42.641
AP College	468	444596	1368258	102.85
AP Dunhuan	63	250782	597289	23.217
AP Jinan	76	1228959	2864740	73.429
AP Vaihingen	20	554169	1201982	6.134
AP Toronto	13	113685	239279	2.523
New College (stereo)	3500	449096	2124449	139.72

*Run on the Virtual Box on an Intel Xeon CPU E5-2690@2.9GHz CPU. Times include building initial reconstructions by BA and the hierarchical process.

For the six AP datasets, the hierarchical BA converged to small mean square re-projection errors as given in Table 3 (with much more computational cost due to the iterations in each BA step). However, for the two Malaga datasets as well as the KITTI dataset, BAs in the hierarchical process do not converge to the correct result for the last step when large gaps are encountered during loop closure (last two steps for the CAMPUS-2L datasets since there are two large loops). This is because when closing the large loop, the initial estimates obtained from the two previous BA results are not good enough. The reason why the proposed linear SFM can successfully close large loops is due to the correct usage of the information matrices which represent the uncertainties of the local reconstructions at each level of fusion, as well as the benefits from solving the linear least squares problem.

The global BA starting from the results of Linear SFM as initial estimates has also been implemented. For all nine datasets, the algorithm converged with

MSEs of less than one pixel^2 and for the six AP datasets, the global BA results are exactly the same as those by hierarchical BA. For the datasets tested, the global BA result starting from Linear SFM is found to have the smallest MSE as compared with global BA results starting from other initial estimates.

The estimated poses from global BA are shown in Figures 4 to 13 in blue lines. The results of the hierarchical BA for PARKING-6L, CAMPUS-2L and KITTI are very poor since the loop is not closed successfully. Thus, they are not shown to save space. The convergence of both hierarchical BA and global BA as well as the computational cost of hierarchical BA are given in Table 3. The relative accuracy (as compared with the global BA) and the computational cost of Linear SFM for different datasets are shown in Tables 1 and 2.

Table 2: RMSE* of Pose Positions by Linear SFM Algorithm

Dataset	Absolute	Relative
PARKING-6L	0.577 m	0.007 m
CAMPUS-2L	4.819 m	0.144 m
KITTI 09	1.242 m	0.035 m
AP Village	5.420e-4	0.712e-4
AP College	7.791e-2	1.064e-2
AP Dunhuan	3.492e-2	0.470e-2
AP Jinan	4.004e-2	0.200e-2
AP Vaihingen	4.970e-4	1.780e-4
AP Toronto	9.265e-4	2.894e-4
New College (stereo)	0.487 m	0.002 m

*All the root mean square errors (RMSE) are given with respect to the results of global BA (to guarantee the convergence, the results of the linear approach are used as the initial estimates in global BA). Relative scales are used in the aerial photogrammetric datasets because of the lack of ground truth.

6.5. Comparison to LinearCamReg

The proposed Linear SFM algorithm is also compared to the linear algorithm proposed in [16] (LinearCamReg) by using AP datasets. To make the inputs to the proposed Linear SFM and LinearCamReg equal, the initial reconstructions from the proposed Linear SFM are used as the triplets in LinearCamReg. The matches for each edge (image pair) are the same as what are used to build the initial reconstruction for the proposed Linear SFM. The relative pose for each edge is given by the BA result for building each initial reconstruction. The accuracy of rotation and translation obtained by the two methods is given in Table 4 and the results of both poses and 3D features of Jinan dataset from the two linear algorithms are shown in Figure 14. As the LinearCamReg algorithm only provides the global poses, the 3D features for LinearCamReg are obtained by triangulation. As we can see from the comparison, the proposed Linear SFM algorithm outperforms LinearCamReg in terms of accuracy of both camera poses and 3D features. This is mainly due to the fact that the information matrices are used at different levels of fusion in our proposed Linear SFM. However, the computational time for LinearCamReg is shorter mainly because it only solves a pose-graph problem and features are not involved in the global camera pose registration.

The Malaga and KITTI datasets are not used in this comparison because not all edges can pass the triplet verification step required in [16]. This is probably due to the fact that the camera is facing and moving forward in some scenarios in these three datasets rather than facing sideways as in the AP datasets.

6.6. Result for Stereo Images

The idea of the proposed linear approach is also applicable to SFM using stereo vision. To ensure the quality of the initial reconstructions, a modified version of ParallaxBA is first used to build the initial reconstructions using stereo images from two camera poses, and then transformed into an XYZ presentation. After building initial reconstructions, the stereo SFM problem can be solved similarly to the Linear monocular SFM as described in Section 4.

Table 3: Comparison to Hierarchical BA [4]

	Hierarchical BA		Global BA	
Dataset	MSE	Time[1]	MSE[2]	Time[3]
PARKING-6L	1.7550e4	296.4	0.0241	29.74+1.984
CAMPUS-2L	3.7864e9	535.3	0.0256	47.69+12.05
KITTI 09	2.5345e2	733.1	0.0101	122.4+6.831
AP Village	0.0716	387.4	0.0716	42.64+4.892
AP College	0.4433	691.6	0.4433	102.9+7.681
AP Dunhuan	0.1682	309.1	0.1682	23.22+3.812
AP Jinan	0.1264	713.7	0.1264	73.43+9.081
AP Vaihingen	0.1193	437.2	0.1193	6.134+6.164
AP Toronto	0.0418	113.5	0.0418	2.523+2.070

1. In seconds, the maximum number of iterations is set to 100 in the Levenberg-Marquardt algorithm.

2. Global BA is started from the result of Linear SFM. The MSE is the minimal among MSEs of global BA started from different initializations.

3. Time is Linear SFM + Global BA.

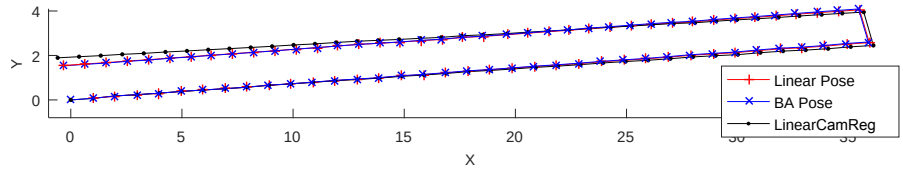
Table 4: Comparison to LinearCamReg [16] w.r.t RMSE[1] of Poses using Aerial Photogrammetric Datasets

	Linear SFM		LinearCamReg	
Dataset	Rotation[2]	Translation[3]	Rotation[2]	Translation[3]
AP Village	0.312e-4	0.054e-2	4.740e-4	1.244e-2
AP College	5.872e-3	0.779e-1	1.606e-3	4.310e-1
AP Dunhuan	1.899e-3	3.492e-2	2.376e-3	5.415e-2
AP Jinan	2.770e-3	0.400e-1	6.643e-3	1.109e-1
AP Vaihingen	0.967e-4	0.050e-2	1.710e-4	1.333e-2

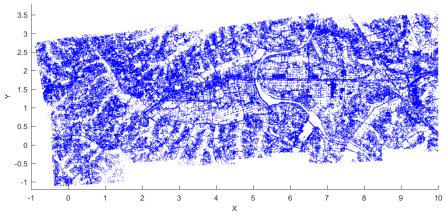
1. All the RMSEs are with respect to the results of global BA.

2. In rad for the rotation angles.

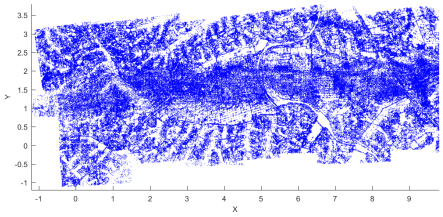
3. Relative scales are used for translations because of the lack of ground truth.



(a) Pose comparison



(b) 3D Reconstruction from Linear SFM



(c) 3D Reconstruction from LinearCamReg [16]

Figure 14: Comparison between the proposed Linear SFM algorithm and the LinearCamReg algorithm in [16] using the photogrammetric Jinan dataset.

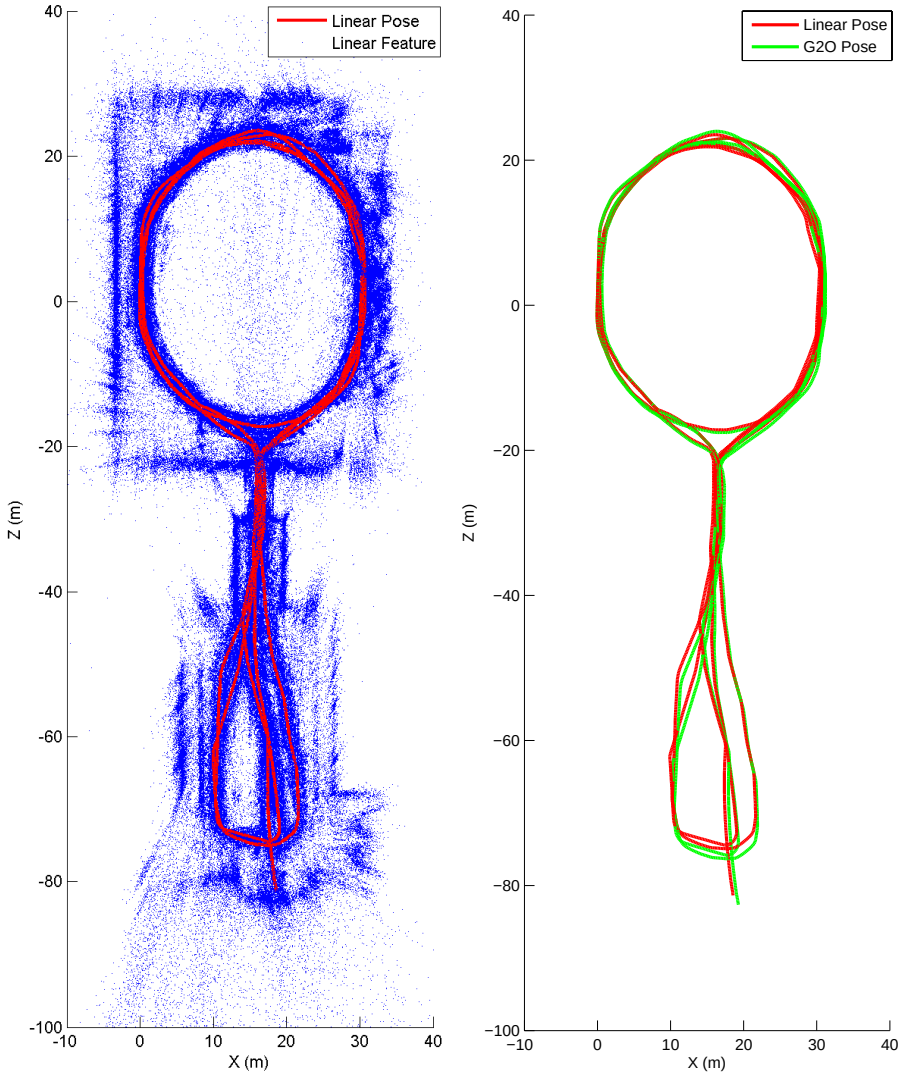
The New College dataset [36] is employed for SLAM with stereo vision. We use the pre-processed dataset from G2O [25] released on OpenSLAM, in which the data association has already been done. There are 3,500 poses, 449,096 features and 2,124,449 projections in total. The result is shown in Figure 15(a). The result is compared with that of global BA (performed using G2O) in Figure 15(b).

The accuracy (as compared with global BA) and the computational cost of Linear SFM using this dataset is summarized in the last row of Tables 1 and 2.

7. Discussion

7.1. The Importance of Information Matrix

The proposed Linear SFM approach only uses linear least squares optimizations plus nonlinear transformations. This overcomes some fundamental limitations of most of the nonlinear optimization based approaches for BA, namely the difficulty of getting a good initial estimate, large number of iterations and



(a) Pose Feature (b) Comparison

Figure 15: Result of New College stereo dataset.

convergence to a local minimum. As compared with existing linear approaches, one key difference is that the information matrices of local reconstructions are used as the weights for the linear least squares optimization (6).

Particularly, after building the initial reconstructions by BA, all the information from original image point observations is summarized in the state estimates and the corresponding information matrices of the initial reconstructions. Later on, this information is fused during the hierarchical joining process in the proposed Linear SFM algorithm. Thus, the final results make use of all the original observation information and there is no information loss in Linear SFM. This makes the results of Linear SFM very close to those from the optimal BA. Thus this approach is fundamentally different from the trajectory registration approaches [16] in which some information is ignored.

The information matrix of a local reconstruction contains the correlation among all the poses and features in the local reconstruction. This correlation is critical in obtaining a good quality estimate in the local reconstruction merging. When joining two local reconstructions, not only the estimates of common features are changed, the estimates of all other poses and features in the local reconstructions are changed accordingly, which means the poses and features are optimized at the same time, rather than optimizing poses first and then optimizing the structure using the obtained pose estimates. If we change the information matrices into identity (without considering any weight in the least squares optimization in (6)) or limit them to be block diagonal (only considering the uncertainties of poses/features, but not considering their correlations), the results obtained will be very poor and similar to the one shown in the upper right subfigure in Figure 6 (before performing linear least squares).

The use of information matrices in the proposed Linear SFM algorithm result in superior performance when closing a large loop as compared with hierarchical BA [4][32]. As the experimental results shown in Section 6.4, in three out of nine datasets, for hierarchical BA, the trajectory drifts too much before a big loop closure such that the final BA cannot correct it with loop closure observations. For Linear SFM, all the loops are closed successfully for all the nine test datasets.

7.2. Size and Quality of Initial Reconstructions

For the proposed Linear SFM algorithm, the initial reconstructions can be of any size (minimal three frames for monocular, two frames for stereo). In this paper, we suggest to build the initial reconstructions with three images simply for minimising the nonlinear part in the whole algorithm.

The quality of the initial reconstructions is important for the proposed Linear SFM algorithm. As only three frames are used in the BA of the suggested initial reconstruction (two frames for stereo case), it is very easy to initialize BA and BA usually converges within few iterations. To ensure the quality of the initial reconstructions, ParallaxBA [22] is used to first build the initial reconstructions and then the features are transformed into an XYZ representation. For all the datasets used in this paper, with initial estimates computed by two view geometry, ParallaxBA using three frames converged in three to five iterations with the mean square of the re-projection errors around 0.1 pixel^2, while BA using XYZ parametrizaton took more iterations to converge and resulted in re-projection errors about twice as large. This is because of the singularity problems involved in BA using XYZ parametrizaton [22].

It should also be mentioned that in the experimental results in this paper, although very far features with near zero parallax appear in many of the initial reconstructions resulting in a large uncertainty in feature XYZ position, they do not have much impact on the final Linear SFM results. This is probably because of the linear least squares based joining approach and the accurate information matrices used. (In our experience, the joining approach using nonlinear optimization has issues with very far features [13].)

It is also interesting to note that although some outliers exist, the initial reconstruction using ParallaxBA has a good quality and the impact of outliers is negligible in our experimental results. Further investigation is necessary to understand how much tolerance on outliers the algorithm has.

7.3. An Approximation to the Globally Optimal BA

Although the Linear SFM results are very good for the practical datasets tested, the proposed linear approach is still an approximation to the (globally optimal) BA.

The difference between the result of Linear SFM and the optimal solution to the nonlinear BA for SFM problems comes from two reasons: (i) Instead of using the original information of image point correspondence observation, we summarize the local reconstruction information as the state estimate together with its uncertainty (information matrix) and use this information in the reconstruction merging process. (ii) Instead of fusing all the reconstructions together in one go using nonlinear optimization, two local reconstructions are fused at each time, resulting in a suboptimal solution.

If the optimal BA result is really desired, the result obtained using the proposed linear approach can serve as an excellent initial estimate for the global BA to get the optimal solution, as shown in Table 3.

8. Conclusion and Future Work

This paper presents an approach to solving SFM problems that mainly uses linear least squares. It is based on the idea of joining local reconstructions and the only nonlinear optimization required is for building the initial reconstructions using BA. The initial reconstructions are joined in a hierarchical manner using only linear least squares and coordinate and scale transformations. The algorithm can be applied to both SFM with monocular camera and SFM with stereo camera. Experimental results demonstrate that the new approach can generate a camera trajectory and feature structure very close to that obtained using global BA.

The reason why linear least squares can be used in the proposed SFM algorithm is that the two local reconstructions to be fused are in the same coordinate frame with the same scale. Thus, the nonlinear component of the optimization problem can be decoupled from the linear component. This is different from the

existing nonlinear optimization based local reconstruction merging algorithms such as [19][12] where the local reconstructions are in different coordinate frames.

Since nonlinear optimization is only used for the building of small size initial reconstructions (containing three camera poses for monocular or two for stereo), good initial estimates of the nonlinear optimization can be obtained easily without worrying about local minima. The joining of these initial reconstructions only requires linear least squares optimization and nonlinear transformations for which no initialization and iterations are needed. Thus, the proposed approach overcomes a fundamental limitation of most of the existing nonlinear optimization based approach for BA, namely the difficulty of getting a good initialization for converging to the global minimum.

The information matrix plays an important role in the proposed Linear SFM algorithm. Linear SFM can successfully close large loops (as shown in the experimental results) due to that we compute and use the information matrices of the local reconstructions correctly at each level of fusion.

In the proposed approach, it is assumed that the images are ordered and taken by calibrated cameras. For images taken from different uncalibrated cameras as in [5], the proposed approach is not appropriate since the geometrical information of a local reconstruction is used as an integrated measurement in the joining process in which the camera intrinsic parameters cannot be optimized.

Future research work includes more efficient implementations of the proposed linear approach, the integration of the proposed approach with robust and efficient feature tracking and matching algorithms, and the extension of the approach to more general SFM problems such as unordered image sequences.

Acknowledgment

The authors would like to thank Dr. Zhaopeng Cui for providing us with the code of LinearCamReg and helping us in its use. The authors would also like to thank the anonymous reviewers for the valuable comments and suggestions on improving the quality of the manuscript.

References

[1] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2000.

[2] N. Snavely, S. M. Seitz, R. Szeliski, Skeletal graphs for efficient structure from motion, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[3] M. I. Lourakis, A. A. Argyros, SBA: A software package for generic sparse bundle adjustment, ACM Transactions on Mathematical Software 36 (1) (2009) 2:1–2:30.

[4] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, P. Sayd, 3D reconstruction of complex structures with bundle adjustment: an incremental approach, in: Proceedings of IEEE International Conference on Robotics and Automation, 2006, pp. 3055–3061.

[5] S. Agarwal, N. Snavely, S. M. Seitz, R. Szeliski, Bundle adjustment in the large, in: Proceedings of European Conference on Computer Vision, 2010, pp. 29–42.

[6] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, R. Szeliski, Building Rome in a day, in: Proceedings of IEEE International Conference on Computer Vision, 2009, pp. 72–79.

[7] Y. Jeong, D. Nistr, D. Steedly, R. Szeliski, I.-S. Kweon, Pushing the envelope of modern methods for bundle adjustment, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (8) (2012) 1605–1617.

[8] Y.-D. Jian, D. C. Balcan, F. Dellaert, Generalized subgraph preconditioners for large-scale bundle adjustment, in: Proceedings of IEEE International Conference on Computer Vision, 2011, pp. 295–302.

[9] F. Kahl, D. Henrion, Globally optimal estimates for geometric reconstruction problems, International Journal of Computer Vision 74 (1) (2007) 3–15.

[10] L. A. Clemente, A. J. Davison, I. Reid, J. Neira, J. D. Tardós, Mapping large loops with a single hand-held camera, in: Proceedings of Robotics: Science and Systems, 2007.

[11] P. Piniés, J. D. Tardós, Large-scale SLAM building conditionally independent local maps: Application to monocular vision, IEEE Transactions on Robotics 24 (5) (2008) 1094–1106.

[12] K. Ni, D. Steedly, F. Dellaert, Out-of-core bundle adjustment for large-scale 3D reconstruction, in: Proceedings of IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[13] L. Zhao, S. Huang, L. Yan, G. Dissanayake, Parallax angle parametrization for monocular SLAM, in: Proceedings of IEEE International Conference on Robotics and Automation, 2011, pp. 3117–3124.

[14] H. Mayer, Efficient hierarchical triplet merging for camera pose estimation, in: Proceedings of German Conference on Pattern Recognition, 2014, pp. 399–409.

[15] L. M. Paz, J. D. Tardós, J. Neira, Divide and conquer: EKF SLAM in O(n), IEEE Transactions on Robotics 24 (5) (2008) 1107–1120.

[16] N. Jiang, Z. Cui, P. Tan, A global linear method for camera pose registration, in: Proceedings of IEEE International Conference on Computer Vision, 2013, pp. 481–488.

[17] L. Zhao, S. Huang, G. Dissanayake, Linear MonoSLAM: A linear approach to large-scale monocular slam problems, in: Proceedings of IEEE International Conference on Robotics and Automation, 2014, pp. 1517–1523.

[18] H.-Y. Shum, Q. Ke, Z. Zhang, Efficient bundle adjustment with virtual key frames: A hierarchical approach to multi-frame structure from motion, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 538–543.

[19] S. Huang, Z. Wang, G. Dissanayake, Sparse local submap joining filter for building large-scale maps, IEEE Transactions on Robotics 24 (5) (2008) 1121–1130.

[20] H. Li, R. Hartley, Five-point motion estimation made easy, in: Proceedings of International Conference on Pattern Recognition, 2006, pp. 630–633.

[21] D. Nistr, An efficient solution to the five-point relative pose problem, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (6) (2004) 756–770.

[22] L. Zhao, S. Huang, Y. Sun, L. Yan, G. Dissanayake, ParallaxBA: Bundle adjustment using parallax angle feature parametrization, International Journal of Robotics Research 34 (4-5) (2015) 493–516.

[23] H. Strasdat, J. Montiel, A. Davison, Scale drift-aware large scale monocular SLAM, in: Proceedings of Robotics: Science and Systems, 2010.

[24] K. Konolige, Sparse sparse bundle adjustment, in: British Machine Vision Conference, 2010, pp. 1–11.

[25] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, W. Burgard, G2O: A general framework for graph optimization, in: Proceedings of IEEE International Conference on Robotics and Automation, 2011, pp. 3607–3613.

[26] C. Wu, S. Agarwal, B. Curless, S. Seitz, Multicore bundle adjustment, Computer Vision and Pattern Recognition 34 (8) (2011) 3057–3064.

[27] L. Zhao, S. Huang, Y. Sun, G. Dissanayake, ParallaxBA: Bundle adjustment using parallax angle feature parametrization, Source code on OpenSLAM.
URL `https://www.openslam.org/`

[28] J.-L. Blanco, F.-A. Moreno, J. Gonzalez, A collection of outdoor robotic datasets with centimeter-accuracy ground truth, Autonomous Robots 27 (4) (2009) 327–351.

[29] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the KITTI vision benchmark suite, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.

[30] M. Cramer, The DGPF-test on digital airborne camera evaluationoverview and test design, Photogrammetrie-Fernerkundung-Geoinformation (2) (2010) 73–82.

[31] F. Rottensteiner, G. Sohn, J. Jung, M. Gerke, C. Baillard, S. Benitez, U. Breitkopf, The ISPRS benchmark on urban object classification and 3D building, ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 1 (3) (2012) 293–298.

[32] E. Royer, M. Lhuillier, M. Dhome, T. Chateau, Localization in urban environments: monocular vision compared to a differential GPS sensor, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 114–121.

[33] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.

[34] M. A. Fischler, 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) (1981) 381–395.

[35] Y. Sun, L. Zhao, S. Huang, L. Yan, G. Dissanayake, L2-SIFT: SIFT feature extraction and matching for large images in large-scale aerial photogrammetry, ISPRS Journal of Photogrammetry and Remote Sensing 91 (2014) 1–16.

[36] M. Smith, I. Baldwin, W. Churchill, R. Paul, P. Newman, The new college vision and laser data set, International Journal of Robotics Research 28 (5) (2009) 595–599.