2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Linear MonoSLAM: A Linear Approach to Large-Scale Monocular SLAM Problems

Liang Zhao, Shoudong Huang and Gamini Dissanayake

Abstract— **This paper presents a linear approach for solving monocular simultaneous localization and mapping (SLAM) problems. The algorithm first builds a sequence of small initial submaps and then joins these submaps together in a divide-and-conquer (D&C) manner. Each of the initial submap is built using three monocular images by bundle adjustment (BA), which is a simple nonlinear optimization problem. Each step in the D&C submap joining is solved by a linear least squares together with a coordinate and scale transformation. Since the only nonlinear part is in the building of the initial submaps, the algorithm makes it possible to solve large-scale monocular SLAM while avoiding issues associated with initialization, iteration, and local minima that are present in most of the nonlinear optimization based algorithms currently used for large-scale monocular SLAM. Experimental results based on publically available datasets are used to demonstrate that the proposed algorithms yields solutions that are very close to those obtained using global BA starting from good initial guess.**

I. Introduction

The monocular simultaneous localization and mapping (MonoSLAM) or structure-from-motion (SFM) problem considered in this paper refers to the process of estimating the three-dimensional structure and the camera trajectory from a sequence of images captured by a single camera [1][2]. Once the association among the features present in a set of images are available, bundle adjustment (BA) can be used to obtain the trajectory and structure by solving a nonlinear least squares problem that minimizes the re-projection errors [3]. In general, for small-scale problems BA easily converges to the globally optimal solution. However, when the number of images is large, BA can be very time-consuming and can converge to a local minimum, unless a good initial guess to the structure and motion is used.

To improve the convergence and efficiency of BA and avoid local minima, skeletal graphs are proposed by Snavely et. al. [4], where a small skeletal subset of images are used to reconstruct the skeletal set, then the remaining leaf images are added using pose estimation. This strategy for improving the quality of initial guess is used in [5][6] for the city-scale 3D reconstruction using conjugate gradient method 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 gradients is proposed in [8]. However, local optimization method for the global BA often requires a large number of iterations to converge [5] even if a good initial value is available.

Building small-scale submaps and then combining the submaps to build the global map is another efficient way to solve large-scale monocular SLAM problems [9][10][11][12]. In [13], a number of submaps are first independently built, then the variables in the submaps that are not directly used in the process of merging of submaps are factored out in order to speed up the submap joining process. In [14], the relative scales between submaps 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 all the above submap based algorithms, the process of combining the submaps requires a solution to a nonlinear optimization problem, thus the initialization is an important issue that require further investigation [13].

This paper presents a linear approach for solving monocular SLAM problems by combining small submaps. The process begins by building a set of initial submaps by BA. Then a large-scale monocular SLAM problem is solved by joining these initial submaps through a divide-and-conquer (D&C) [15] process. An initial submap is defined as a small submap built using three images. Building these initial submaps is the only part requiring nonlinear optimization. Each step of joining two submaps in the D&C process is formulated as a linear least squares problem by judiciously selecting an appropriate coordinate and scale transformation of the submaps. Evaluations using publicly available datasets show that the solutions obtained using the linear approach are very accurate. If really necessary, a global BA can be performed using the Linear MonoSLAM result as the initial value.

Work presented in this paper is based on our recent work on Linear SLAM [16] where range and bearing sensors are required. The major additional challenge in Linear MonoSLAM is that the different initial submaps built using three monocular images have different scales. Thus a new strategy for a scale and coordinate transformation is required before the two submaps can be joined in a linear way.

The paper is organized as follows. Section II provides some preliminaries for submap joining approach. Section III presents the framework of the proposed linear algorithm for monocular SLAM. Some details of the Linear MonoSLAM algorithm is presented in Section IV. Section V presents the experimental results using large-scale datasets. Finally, Section VI concludes the paper.

This work is supported by Australian Research Council (ARC) Discovery grant DP120102786. L. Zhao, S. Huang and G. Dissanayake are with the Centre for Autonomous Systems, Faculty of Engineering and IT, University of Technology, Sydney, NSW2007, Australia. {Liang.Zhao-1, Shoudong.Huang,Gamini.Dissanayake}@uts.edu.au

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

II. Preliminaries for Submap Joining

A. Submap Joining in SLAM

It is well known that the problem of point feature based range-bearing SLAM [17] can be formulated as a nonlinear optimization problem

$$\text{minimize} \quad \|Z - f(\mathbf{X})\|^2_{I_Z} \tag{1}$$

where Z is the vector of measurements (observations and odometry information), $\mathbf{X}$ is the state vector containing all the feature positions and the robot poses, $f(\mathbf{X})$ is the nonlinear function describing the relation between $\mathbf{X}$ and Z, and I_Z is the information matrix (inverse of the covariance matrix) of the measurement noises.

Submap joining has shown to be an efficient strategy for solving large-scale SLAM problems. The idea of submap joining is to separate the measurements Z into different parts and use each part to build a small submap, then combine the submaps to get the global map. Suppose the measurements are divided into two parts and each part is used to build one submap. The two submaps are denoted by $(\hat{\mathbf{X}}_1, I_{X_1})$ and $(\hat{\mathbf{X}}_2, I_{X_2})$. Here $\mathbf{X}_1$ and $\mathbf{X}_2$ are the state vectors of the two submaps, defined in two different local coordinate frames. I_{X_1} and I_{X_2} are the corresponding information matrices.

When combining the two submaps, one can use the submaps $(\hat{\mathbf{X}}_1, I_{X_1})$ and $(\hat{\mathbf{X}}_2, I_{X_2})$ as two integrated measurements to build the global map $(\hat{\mathbf{X}}, I_X)$ [18]. Since the information matrices I_{X_1} and I_{X_2} represent the uncertainty of the submap estimates $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$, they are used as the weights of the nonlinear least squares problem [19]. Thus the optimization problem becomes

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

where $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$ are the functions relating the global state vector $\mathbf{X}$ to the submap state vectors.

In traditional submap joining, the submap state vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are defined in different coordinate frames, at least one of the functions $f_1(\mathbf{X})$ and $f_2(\mathbf{X})$ is nonlinear no matter which coordinate frame is used for defining the global state vector $\mathbf{X}$. Thus the submap joining problem (2) is a nonlinear optimization problem.

B. Submap Joining in Linear SLAM

Different from the traditional submap joining algorithms as described in Section II-A, in Linear SLAM [16], submap $(\hat{\mathbf{X}}_1, I_{X_1})$ is built in the coordinate frame defined by its end pose, while submap $(\hat{\mathbf{X}}_2, I_{X_2})$ is built in the coordinate frame defined by its start pose, which is the same as the end pose of $(\hat{\mathbf{X}}_1, I_{X_1})$. The coordinate frame of the global map $(\hat{\mathbf{X}}, I_X)$, is defined by the robot end pose of submap $(\hat{\mathbf{X}}_2, I_{X_2})$, or the robot start pose of submap $(\hat{\mathbf{X}}_1, I_{X_1})$.

It is shown in [16] that although the joining of the two submaps to get the global map in (2) is still a nonlinear optimization problem, it is equivalent to building the global map in the coordinate frame defined by the end pose of the first submap by a linear least squares optimization, plus a nonlinear coordinate transformation. Please refer to [16] for the details.

C. Submap Joining in Monocular SLAM

The submap joining idea can also be applied to monocular SLAM problem. However, for monocular SLAM, the absolute scale cannot be observed by a single camera unless some external information is available. Thus the scales in different submaps will be different. When joining the submaps with different scales, the observation function $f_1(\cdot)$ and $f_2(\cdot)$ in (2) must be carefully formulated by considering the relative scale between the two submaps.

In the next sections, we will show that similar to Linear SLAM for joining submaps built from range-bearing information, submap joining in monocular SLAM can also be formulated as a linear least squares problem by carefully selecting the coordinate frames and performing coordinate and scale transformation.

III. The Framework of Linear MonoSLAM

A. Building Initial Submaps

In the proposed linear algorithm for monocular SLAM, the only nonlinear optimization part is the building of a sequence of small submaps for the linear submap joining algorithm. Thus, we propose to build these submaps as small as possible and call them initial submaps. This will make the whole process of solving monocular SLAM problem as linear as possible. In this paper, each initial submap is built with 3 images, and there are two common camera poses between two adjacent initial submaps. The reason for having two common poses, instead of one common pose as in the submap joining algorithms in the traditional range-bearing SLAM, is to make sure that the relative scale between two adjacent submaps can be worked out easily.

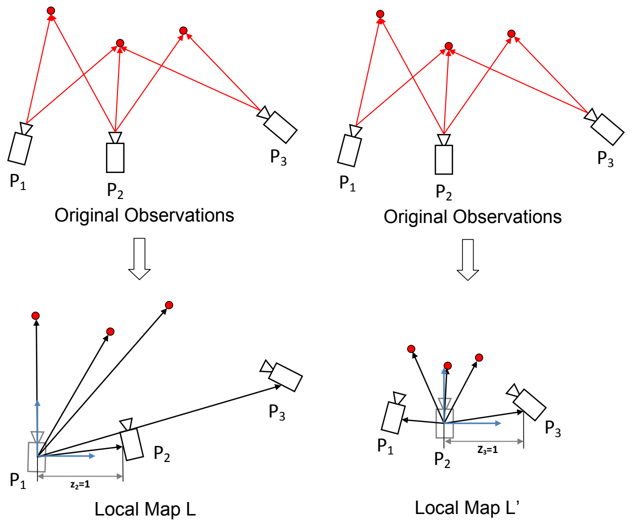
In order to get the best quality of the initial submaps, BA is used to build the initial submaps. BA is the gold standard for monocular SLAM as it solves the optimization problem involving all observations as shown in Fig. 1.

When performing BA, 7 degree of freedom (DoF), namely 6 DoF for coordinate frame and 1 DoF for scale, should be fixed [20]. 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. Without loss of generality, we assume the translation in the Z direction is the largest element in the translation vector. Thus we can fix the z value of the translation from one pose to another pose as 1 to define the scale.

B. Submap Transformation

When building submaps in Section III-A, if we fix the first pose $\mathbf{P}_1$ as $\mathbf{0}$ to define the coordinates frame, and fix the z value of the translation from the first pose $\mathbf{P}_1$ to the second pose $\mathbf{P}_2$ as 1 to define the scale of the submap, then we can get submap $L = (\hat{\mathbf{X}}, I)$ in Fig. 1(a). If we fix the second pose $\mathbf{P}_2$ as $\mathbf{0}$ to define the coordinates frame, and fix the z value of the translation from the second pose $\mathbf{P}_2$ to the third pose $\mathbf{P}_3$ as 1 to define the scale of the submap, then we can get submap $L' = (\hat{\mathbf{X}}', I')$ in Fig. 1(b).

Note that the two submaps in Fig. 1, L and L', can be easily converted from one to another by applying coordinate



(a) Building Submap L (b) Building Submap L'

Fig. 1. Building different submaps from the same observations by using different coordinate frames and different scales. The two submaps can be transformed to each other by applying a coordinate and scale transformation.

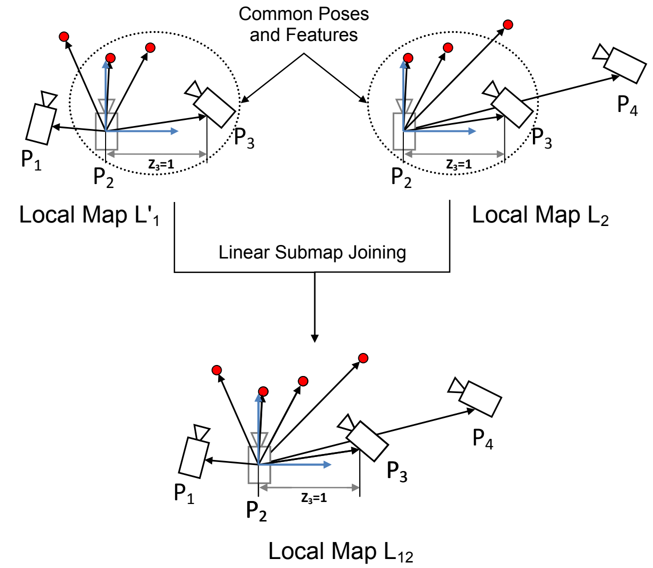


Fig. 2. When the two submaps are in the same coordinate frame with the same scale, joining the two submaps becomes a linear problem.

and scale transformation. Thus there are two ways to obtain L'. One is to solve BA by defining $\mathbf{X}'$ as the state vector, another is to first get the optimal estimate of the submap $L = (\hat{\mathbf{X}}, I)$, then apply a coordinate and scale transformation. The results using these two ways are identical provided that the scale is not degenerated by zero translation (The degenerate case can be avoided by simply selecting the keyframes.)

The key idea of our Linear MonoSLAM algorithm is to make necessary coordinate and scale transformation on the submaps such that the two submaps to be joined together are in the same coordinate frame and with the same scale.

C. Joining Two Submaps

Suppose the two submaps to be joined together are L'_1 and L_2 as shown in Fig. 2. Here L'_1 is built by using the projections from Image 1, 2 and 3, and L_2 is built by using the projections from Image 2, 3 and 4.

Both L'_1 and L_2 are in the coordinate frame of $\mathbf{P}_2$, with the z value of the translation from $\mathbf{P}_2$ to $\mathbf{P}_3$ equal to 1 as scale. Thus the two submaps are in the same coordinates and with the same scale. So we can define the global state vector $\mathbf{X}$ in the same coordinate frame, and the joining of these two submaps can be solved as a linear least squares problem, as seen in Fig. 2.

D. Solving the MonoSLAM Problem by Divide-and-Conquer

The process of joining a sequence of submaps using divide-and-conquer (D&C) method is illustrated in Fig. 3. The structure is similar to that in [15]. It can be seen that at each step, only two submaps are joined together. As the level in D&C increases, the size of the two submaps to be joined together becomes larger and larger.

As can be seen from Fig. 3, two submaps L'_1 and L_2 are in the same coordinates with the same scale and they are joined together to build submap L_{12}. Similarly, two submaps L'_3 and L_4 are also in the same coordinates with the same scale and they are joined together to build submap L_{34}. The result of submap L_{12} and submap L_{34} are then transformed into L'_{12} and L'_{34} with the same coordinate frame $\mathbf{P}_3$. The two submaps L'_{12} and L'_{34} are then joined together to get submap L_{1234}.

Since the two submaps to be joined together are always in the same coordinates with the same scale, the joining process is a linear least squares problem. Thus, the whole submap joining process can be done by joining a number of initial submaps (e.g. submaps L_1, L_2, L_3, and L_4 in Fig. 3) to build the global map by a D&C method, with only solving linear least squares problems and performing coordinate and scale transformations.

IV. Some Details of Linear MonoSLAM

A. Building Initial Submaps by BA

The original observations of monocular SLAM are the feature projections in the images. Thus the feature projections from three images are used to build an initial submap by estimating the camera poses and feature positions using these observations. Since only three camera poses are involved, the initial value can be easily obtained, such as using the five-point algorithm [21]. After obtaining the optimal estimate of the state vector in BA, the corresponding information matrix can be obtained by $I = J^T J$ where J is the Jacobian of all the projections in BA, evaluated at the optimal estimate of the state vector.

B. Transformation of the Submap

Suppose submap L_1 is given by

$$L_1 = (\hat{\mathbf{X}}_1,\ I_{X_1}) \tag{3}$$

where $\hat{\mathbf{X}}_1$ is the estimate of the state vector $\mathbf{X}_1$, and I_{X_1} is the associated information matrix.

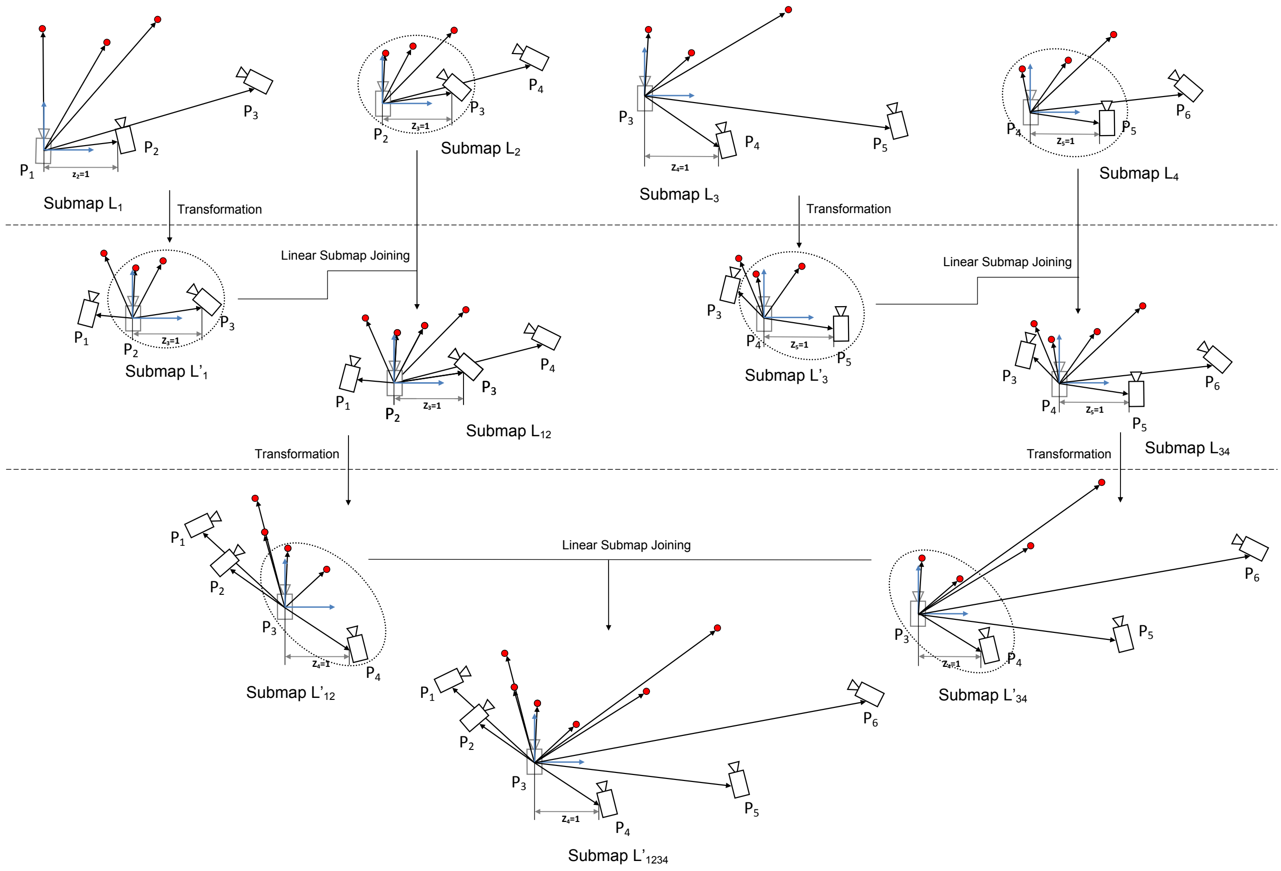


Fig. 3. The proposed divide-and-conquer process in Linear MonoSLAM. The poses and features in the circles are the common poses and features between two submaps.

The state vector $\mathbf{X}_1$ is defined as (for simplicities, some transposes of vectors are omitted in this paper)

$$\mathbf{X}_1 = [{}^1\mathbf{r}_2,\ {}^1x_2,\ {}^1y_2,\ {}^1\mathbf{r}_3,\ {}^1\mathbf{t}_3,\ {}^1\mathbf{F}]. \tag{4}$$

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 the state vector $\mathbf{X}_1$ in (4), pose $\mathbf{P}_3$ in the coordinate frame of $\mathbf{P}_1$ is presented by

$${}^1\mathbf{P}_3 = [{}^1\mathbf{r}_3,\ {}^1\mathbf{t}_3] \tag{5}$$

where ${}^1\mathbf{r}_3 = [{}^1\alpha_3\ {}^1\beta_3\ {}^1\gamma_3]^T$ is the vector containing the three Euler angles, and ${}^1\mathbf{t}_3 = [{}^1x_3,\ {}^1y_3,\ {}^1z_3]^T$ is the translation; ${}^1\mathbf{F}$ represents all the feature XYZ positions in submap L_1. $\mathbf{P}_1 = \mathbf{0}$ is fixed as the coordinate frame and ${}^1z_2 = 1$ is fixed as the scale of submap L_1, thus they are not in the state vector.

The state vector of submap $L_1' = (\hat{\mathbf{X}}_1',\ I_{X_1'})$ is denoted as

$$\mathbf{X}_1' = [{}^2\mathbf{r}_1,\ {}^2\mathbf{t}_1,\ {}^2\mathbf{r}_3,\ {}^2x_3,\ {}^2y_3,\ {}^2\mathbf{F}]. \tag{6}$$

Here $\mathbf{P}_2 = \mathbf{0}$ is fixed as the coordinate frame and ${}^2z_3 = 1$ is fixed as the scale of submap L_1', thus they are not in the state vector $\mathbf{X}_1'$.

The relation between $\mathbf{X}_1$ and $\mathbf{X}_1'$ is the coordinate and scale transformation function given by

$$\mathbf{X}_1' = g(\mathbf{X}_1) \Rightarrow \begin{cases} {}^2\mathbf{r}_1 = r^{-1}({}^1R_2^T) \\ {}^2\mathbf{t}_1 = -{}^1R_2\,[{}^1x_2, {}^1y_2, 1]^T/z_s \\ {}^2\mathbf{r}_3 = r^{-1}({}^1R_3{}^1R_2^T) \\ {}^2x_3 = x_s/z_s \\ {}^2y_3 = y_s/z_s \\ {}^2\mathbf{F} = {}^1R_2\,({}^1\mathbf{F} - [{}^1x_2,{}^1y_2, 1]^T)/z_s \end{cases} \tag{7}$$

where ${}^1R_2 = r({}^1\mathbf{r}_2)$, ${}^1R_3 = r({}^1\mathbf{r}_3)$ are the rotation matrices of pose ${}^1\mathbf{P}_2$ and pose ${}^1\mathbf{P}_3$ in the state vector $\mathbf{X}_1$. And $r(\cdot)$ and $r^{-1}(\cdot)$ are the angle-to-matrix and matrix-to-angle functions.

In (7), the scale factor z_s as well as x_s and y_s can be computed as

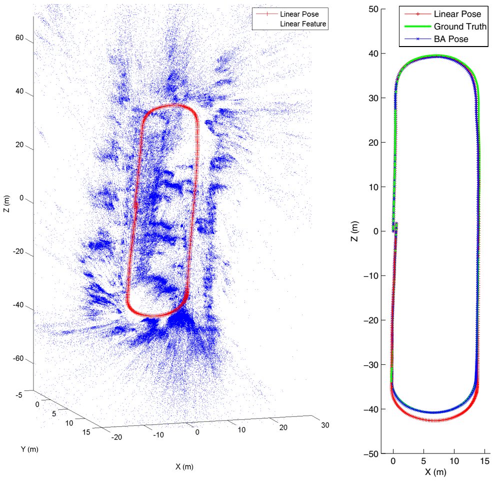
$$[x_s,\ y_s,\ z_s]^T = {}^1R_2\,({}^1\mathbf{t}_3 - [{}^1x_2,{}^1y_2, 1]^T). \tag{8}$$

If we have already got the submap $L_1 = (\hat{\mathbf{X}}_1,\ I_{X_1})$, then the estimate of the state vector $\mathbf{X}_1'$ in submap L_1' can be obtained by

$$\hat{\mathbf{X}}_1' = g(\hat{\mathbf{X}}_1). \tag{9}$$

The corresponding information matrix $I_{X_1'}$ can also be obtained by

$$I_{X_1'} = \nabla^T I_{X_1} \nabla \tag{10}$$



(a) Pose Feature (b) Comparison

Fig. 4. Linear MonoSLAM result of PARKING-6L dataset.

where ∇ is the Jacobian of $\mathbf{X}_1$ with respect to $\mathbf{X}_1'$, evaluated at $\hat{\mathbf{X}}_1'$

$$\nabla = \frac{\partial g^{-1}(\mathbf{X}_1')}{\partial \mathbf{X}_1'}|_{\hat{\mathbf{X}}_1'} \tag{11}$$

Here $\mathbf{X}_1 = g^{-1}(\mathbf{X}_1')$ is the inverse function of $g(\cdot)$ in (7).

C. Joining Two Submaps as a Linear Least Squares Problem

Suppose there are two submaps L_1' and L_2 given by

$$L_1' = (\hat{\mathbf{X}}_1', \ I_{X_1'}), \quad L_2 = (\hat{\mathbf{X}}_2, \ I_{X_2}) \tag{12}$$

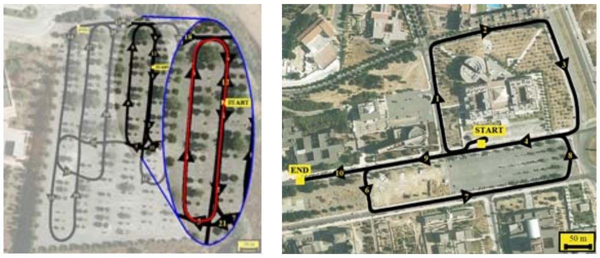
where the state vectors $\mathbf{X}_1'$ and $\mathbf{X}_2$ of submaps L_1' and L_2 are defined as

$$\begin{aligned} \mathbf{X}_1' &= [{}^2\mathbf{r}_1, \ {}^2\mathbf{t}_1, \ {}^2\mathbf{F}_1, \ {}^2\mathbf{r}_3, \ {}^2x_3, \ {}^2y_3, \ {}^2\mathbf{F}_C] \\ \mathbf{X}_2 &= [{}^2\mathbf{r}_3, \ {}^2x_3, \ {}^2y_3, \ {}^2\mathbf{F}_C, \ {}^2\mathbf{r}_4, \ {}^2\mathbf{t}_4, \ {}^2\mathbf{F}_2] \end{aligned} \tag{13}$$

and $I_{X_1'}$ and I_{X_2} are the associated information matrices.

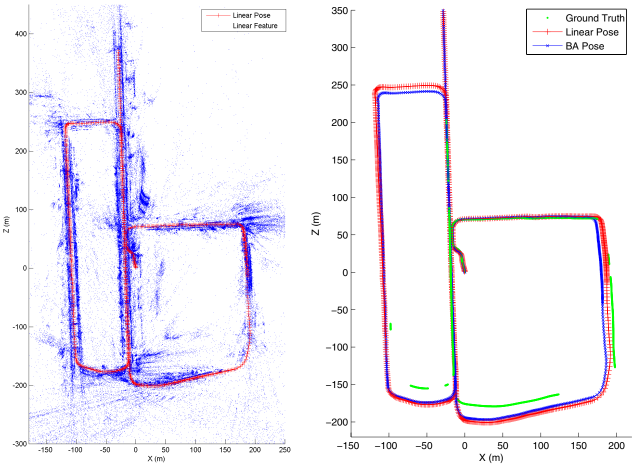
Here $\mathbf{P}_2 = \mathbf{0}$ is fixed as the coordinate frame and ${}^2z_3 = 1$ is fixed as the scale for both L_1' and L_2, thus L_1' and L_2 are in the same coordinate frame, with the same scale.

Instead of using ${}^2\mathbf{F}$ to represent features as in (6), in the state vectors $\mathbf{X}_1'$ and $\mathbf{X}_2$ in (13), ${}^2\mathbf{F}_1, {}^2\mathbf{F}_C, {}^2\mathbf{F}_2$ are used to represent the features, where ${}^2\mathbf{F}_C$ represents the common features appear in both of the two submaps, while ${}^2\mathbf{F}_1$ and ${}^2\mathbf{F}_2$ represent the features only appear in L_1' or L_2, respectively.



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

Fig. 5. The trajectories of Malaga monocular datasets.



(a) Pose Feature (b) Comparison

Fig. 6. Linear MonoSLAM result of CAMPUS-2L dataset.

Denote the state vector of submap L_{12} as

$$\mathbf{X} = [{}^2\mathbf{r}_1, \ {}^2\mathbf{t}_1, \ {}^2\mathbf{F}_1, \ {}^2\mathbf{r}_3, \ {}^2x_3, \ {}^2y_3, \ {}^2\mathbf{F}_C, \ {}^2\mathbf{r}_4, \ {}^2\mathbf{t}_4, \ {}^2\mathbf{F}_2]. \tag{14}$$

Here all the variables are in the coordinate frame of $\mathbf{P}_2$, with ${}^2z_3 = 1$ as the scale, so $\mathbf{P}_2$ and ${}^2z_3$ are not included in $\mathbf{X}$.

Because the two submaps L_1' and L_2, as well as the submap L_{12} are all in the same coordinates with the same scale, the observation functions $f_1(\cdot)$ and $f_2(\cdot)$ in (2) of submap joining becomes linear, thus the joining of two submaps L_1' and L_2 to build the submap $L_{12} = (\hat{\mathbf{X}}, \ I_X)$ becomes a linear least squares problem

$$\text{minimize} \quad \|\hat{\mathbf{X}}_1' - A_1\mathbf{X}\|^2_{I_{X_1'}} + \|\hat{\mathbf{X}}_2 - A_2\mathbf{X}\|^2_{I_{X_2}} \tag{15}$$

where the coefficient matrices A_1 and A_2 are formed by identity and zero matrices as follows

$$A_1 = \left[\mathrm{I}_{n_1} \mid 0_{n_1 \times m_2}\right], \quad A_2 = \left[0_{n_2 \times m_1} \mid \mathrm{I}_{n_2}\right]. \tag{16}$$

Here n_1 and n_2 are the dimensions of the state vectors $\mathbf{X}_1'$ and $\mathbf{X}_2$ of the two submaps, respectively. $m_1 = 3k_1 + 6$ and $m_2 = 3k_2 + 6$, where k_1 and k_2 are the number of features in ${}^2\mathbf{F}_1$ and ${}^2\mathbf{F}_2$, respectively. Thus m_1 and m_2 are the dimensions of the poses and features in the state vectors $\mathbf{X}_1'$ and $\mathbf{X}_2$, which only appear in L_1' or L_2.

The optimal solution $\hat{\mathbf{X}}$ of this linear least squares problem can be computed by solving the linear equation

$$(A^T I_Z A)\hat{\mathbf{X}} = A^T I_Z \mathbf{Z}. \tag{17}$$

where $\mathbf{Z} = [\hat{\mathbf{X}}_1'^T, \ \hat{\mathbf{X}}_2^T]^T$, $I_Z = diag(I_{X_1'}, I_{X_2})$, and $A = [A_1^T, \ A_2^T]^T$.

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

$$I_X = A^T I_Z A \tag{18}$$

It is obvious that the above linear least squares formulation can be extended to the joining of two submaps with any size as long as they are in the same coordinate frame and with the same scale.

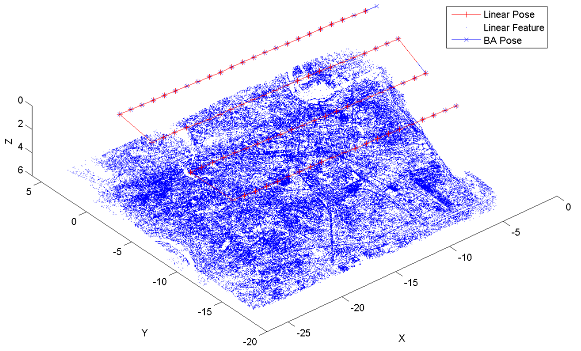


Fig. 7. Linear MonoSLAM result of photogrammetric Village dataset.

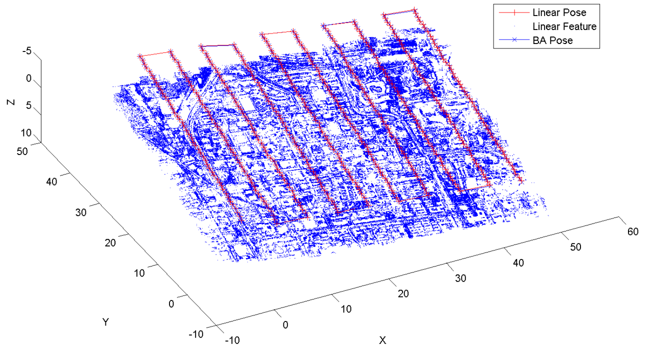


Fig. 8. Linear MonoSLAM result of photogrammetric College dataset.

V. Experimental Results

First publicity available Malaga 2009 Robotic Dataset Collection [22] is used for evaluating the proposed algorithm. As described in [22], a centimeter-level ground truth is provided which can allow us to compare the results of the proposed linear approach.

Images captured by the right camera are used. The image resolution is 1024×768 and camera calibration parameters are provided in the dataset. SIFT [23] and RANSAC [24] are used for feature detection with subpixel accuracy and for matching, including the loop closure matching.

As described in Section III-A, every three images are used to build submaps as the initial submaps by BA. To insure the quality of the initial submaps, in this paper, the initial submaps are first built using BA with parallax angle feature parametrization (ParallaxBA) [14], which has better convergence and accuracy as compared with BA using Euclidean XYZ feature parametrization, and then transformed into XYZ presentation. With initial guess computed by two-view geometry, ParallaxBA using 3 frames converged easily with the mean square of the re-projection errors around 0.1 within 3-5 iterations, while BA using XYZ took more iterations to converge and resulted in re-projection errors about twice as large (because of the singularity problems).

A. PARKING-6L Dataset

First we select one sequence of images collected from a 250m close loop trajectory (Fig. 5(a)) with 508 images. There are 508 poses, 190,711 features and 567,836 projections in total after SIFT matching and RANSAC outlier removal. The result of Linear MonoSLAM is shown in Fig. 4(a). The estimated poses are compared with the result of global BA as well as the ground truth in Fig. 4(b).

B. CAMPUS-2L Dataset

In the CAMPUS-2L dataset (Fig. 5(b)), the 2.2km long trajectory with two loops is used in the experiments. The 1,020 keyframers are selected from the 5,103 images, by simply selecting one from every 5 images. There are 1,020 poses, 198,563 features and 575,644 projections in total by using SIFT and RANSAC. The result of the proposed Linear MonoSLAM approach, the global BA result and the ground truth are shown in Fig. 6.

Two aerial photogrammetric datasets are also used for evaluating the proposed Linear MonoSLAM algorithm. For these two datasets, the cameras are mounted on the aerial plane platforms to map the ground surface.

C. Aerial Photogrammetric Village Dataset

There are 90 images in the Village dataset, which are taken by digital mapping camera (DMC) in snake track with image resolution 7680×13824 pixels. After SIFT and RANSAC are processed, 273,131 features and 779,268 projections are extracted and matched as the input to the proposed Linear MonoSLAM algorithm. The mapping result as well as the camera poses by Linear MonoSLAM are shown in Fig. 7. As comparison, the camera poses result by global BA is also shown in Fig. 7.

D. Aerial Photogrammetric College Dataset

In the College dataset, 468 images with resolution 5616×3744 are captured by Cannon camera. For this dataset, 444,596 features and 1,368,258 projections are obtained after the process of SIFT and RANSAC, and used in the proposed Linear algorithm. The Linear MonoSLAM result as well as the poses of the global BA result are shown in Fig. 8.

The associated video presents the divide-and-conquer process of Linear MonoSLAM for the CAMPUS-2L and College datasets.

VI. Discussion, Conclusion and Future Work

This paper presents a linear approach for solving monocular SLAM problems. The initial submaps are built by BA and then joined together in a divide-and-conquer manner by solving linear least squares problems and applying coordinate and scale transformations. The reason why linear least squares can be used is that the two submaps are transformed into the same coordinate frame with the same scale before they are fused together. Experimental results demonstrated that the linear approach can generate the camera trajectory and feature structure very close to that using global BA.

Since nonlinear optimization is only used for the building of small size initial submaps containing 3 camera poses, good initial value of the nonlinear optimization can be obtained easily without worrying about the local minima issue. The joining of these initial submaps only requires linear least

TABLE I

RMSE* of Pose Positions by Linear MonoSLAM Algorithm

Dataset	Absolute	Relative
PARKING-6L	0.57684567 m	0.00725283 m
CAMPUS-2L	4.81920339 m	0.14385726 m
AP Village	0.00054203	0.00007121
AP College	0.07791178	0.01064497

*All the RMSEs are respect to the results of global BA (to guarantee the convergence, the results of the linear approach are used as the initial guess in global BA). The relative scales are used in the aerial photogrammetric Village and College datasets because of the lack of ground truth.

TABLE II

Computational Costs* of Linear MonoSLAM Algorithm (in seconds)

Dataset	Pose	Feature	Projection	time
PARKING-6L	508	190711	567836	29.736
CAMPUS-2L	1020	198563	575644	47.688
AP Village	90	273131	779268	42.641
AP College	468	444596	1368258	102.854

*Run on the Virtual Box on an Intel Xeon CPU E5-2690@2.9GHz CPU. Times for building initial submaps by BA and the D&C process are both included. Times for data association are not included.

squares thus initialization and iterations are not needed and local minimum does not exist. Thus the proposed approach overcomes a fundamental limitation of most of the existing nonlinear optimization based approach for BA, namely the difficulty of getting good initialization and converging to the global minimum.

The quality of the initial submaps is important to the proposed Linear MonoSLAM algorithm. Thus ParallaxBA is used for building initial submaps which are then transformed into XYZ presentation. In the experimental results in this paper, although very far features with near zero parallax appear in many of the initial submaps resulting in large uncertainty in feature position, they do not have much impact on the final Linear MonoSLAM results probably because of the linear map joining approach (from our experience, submap joining using nonlinear optimization has issues with very far features).

The proposed linear approach is still an approximation to the global BA. If the optimal BA result is really desired, the result obtained using the proposed linear approach can be served as an excellent initial value for the global BA to get the optimal solution.

In the proposed approach, it is assumed that the data association is done (including loop closure data association), it is also assumed that the images are ordered and taken by calibrated cameras with nonzero translation between two consecutive camera poses. Future research work include the integration of the proposed approach with robust and efficient feature tracking and matching algorithms to make it work online, and the extension of the approach to more general visual SLAM problems such as the cases when different uncalibrated cameras are used.

References

[1] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: Real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[2] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[3] M. I. Lourakis and A. A. Argyros, "SBA: A software package for generic sparse bundle adjustment," *ACM Transactions on Mathematical Software*, vol. 36, no. 1, p. 2, 2009.

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

[5] S. Agarwal, N. Snavely, S. M. Seitz, and 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, and R. Szeliski, "Building rome in a day," in *Proceedings of IEEE International Conference on Computer Vision*, 2009, pp. 72–79.

[7] Y. Jeong, D. Nister, D. Steedly, R. Szeliski, and I.-S. Kweon, "Pushing the envelope of modern methods for bundle adjustment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1474–1481.

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

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

[10] C. Estrada, J. Neira, and J. D. Tardós, "Hierarchical SLAM: real-time accurate mapping of large environments," *IEEE Transactions on Robotics*, vol. 21, no. 4, pp. 588–596, 2005.

[11] P. Piniés and J. D. Tardós, "Large-scale SLAM building conditionally independent local maps: Application to monocular vision," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1094–1106, 2008.

[12] G. Grisetti, R. Kummerle, and K. Ni, "Robust optimization of factor graphs by using condensed measurements," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 581–588.

[13] K. Ni, D. Steedly, and 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.

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

[15] L. M. Paz, J. D. Tardós, and J. Neira, "Divide and conquer: EKF SLAM in O(n)," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1107–1120, 2008.

[16] L. Zhao, S. Huang, and G. Dissanayake, "Linear SLAM: A linear solution to the feature-based and pose graph SLAM based on submap joining," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 24–30.

[17] F. Dellaert and M. Kaess, "Square root SAM: Simultaneous localization and mapping via square root information smoothing," *International Journal of Robotics Research*, vol. 25, no. 12, pp. 1181–1203, 2006.

[18] S. Huang, Z. Wang, and G. Dissanayake, "Sparse local submap joining filter for building large-scale maps," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1121–1130, 2008.

[19] S. Huang, Z. Wang, G. Dissanayake, and U. Frese, "Iterated D-SLAM map joining: evaluating its performance in terms of consistency, accuracy and efficiency," *Autonomous Robots*, vol. 27, no. 4, pp. 409–429, 2009.

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

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

[22] J.-L. Blanco, F.-A. Moreno, and J. Gonzalez, "A collection of outdoor robotic datasets with centimeter-accuracy ground truth," *Autonomous Robots*, vol. 27, no. 4, pp. 327–351, 2009.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 9 1–110, 2004.

[24] 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*, vol. 24, no. 6, pp. 381–395, 1981.