Robust Large Scale Monocular Visual SLAM

Guillaume Bourmaud Rémi Mégret
Univ. Bordeaux, CNRS, IMS, UMR 5218, F-33400 Talence, France
{guillaume.bourmaud,remi.megret}@ims-bordeaux.fr

Abstract

This paper deals with the trajectory estimation of a monocular calibrated camera evolving in a large unknown environment, also known as monocular visual simultaneous localization and mapping. The contribution of this paper is threefold: 1) We develop a new formalism that builds upon the so called ***Known Rotation Problem*** *to robustly estimate submaps (parts of the camera trajectory and the unknown environment). 2) In order to obtain a globally consistent map (up to a scale factor), we propose a novel* ***loopy belief propagation*** *algorithm that is able to efficiently align a large number of submaps. Our approach builds a graph of relative 3D similarities (computed between the submaps) and estimates the global 3D similarities by passing messages through a super graph until convergence. 3) To render the whole framework more robust, we also propose a simple and efficient* ***outlier removal*** *algorithm that detects outliers in the graph of relative 3D similarities. We extensively demonstrate, on the TUM and KITTI benchmarks as well as on other challenging video sequences, that the proposed method outperforms the state of the art algorithms.*

1. Introduction

Estimating a 3D model of the environment in which a camera evolves as well as its trajectory, also known as Visual Simultaneous Localization And Mapping (VSLAM), is an important problem for the computer vision community. Indeed, a large number of applications, such as image-based localization [1, 2] or augmented reality, assume that a 3D model of the environment has been previously reconstructed. Thus, being able to accurately estimate this 3D model is essential in order for these applications to operate correctly.

Robust, accurate and scalable VSLAM algorithms for a stereo camera have been proposed [3, 4] a few years ago. However, stereo cameras are still not widely spread compared to monocular cameras which are present on every smart-phone. As a consequence, this paper focuses on the monocular VSLAM problem.

In this problem, one of the major difficulties, compared to stereo VSLAM, consists in the fact that the scale of the scene is not observed. In order to prevent scale drift, loop closures (i.e when the camera comes back at a place already visited) need to be detected. However, a large environment usually contains places that look alike. Thus when a camera evolves in such an environment, wrong loop closures may be detected, resulting in an erroneous 3D model.

We propose a novel robust monocular VSLAM algorithm which is able to operate on long challenging videos where the state of the art algorithms fail. First of all, submaps (parts of the camera trajectory and the unknown environment) are robustly and accurately estimated using the so-called *Known Rotation Problem* [5]. We then build a graph of relative 3D similarities (computed between the submaps). In order to reject the outlier relative 3D similarities coming from wrong loop closures, we propose a simple and efficient *outlier removal algorithm*. Finally, to obtain a scalable monocular VSLAM framework, we derive a *loopy belief propagation* algorithm which is able to align a large number of submaps very efficiently.

The rest of the paper is organized as follows: section 2 deals with the related works. Our novel monocular VSLAM framework is presented in section 3. The two proposed algorithms dedicated to outlier rejection and inference in the graph of relative 3D similarities are described in section 4. In section 5, the limitations of the proposed approach are discussed while in section 6, our monocular VSLAM formalism is evaluated experimentally. Finally, a conclusion is provided in section 7.

2. Related Work

The problem of monocular VSLAM has been studied for 20 years. Thus an exhaustive state of the art is beyond the scope of this paper. Here, we simply describe the most recent approaches and their differences with our novel method. Almost all the recent approaches, as well as the one we propose in this paper, consist in two main modules:

1) A Visual Odometry (VO) approach which estimates the camera poses and the 3D model associated to several

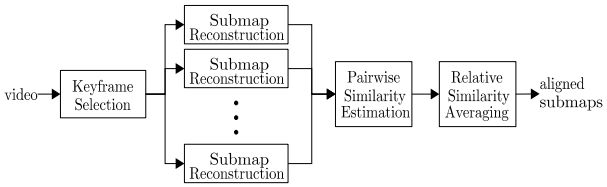


Figure 1: Our Monocular VSLAM Framework

consecutive video frames. In [6] and [7, 8], VO consists in building submaps using a Kalman filter and incremental like Bundle Adjustment (BA), respectively. [9] does not explicitly builds submaps but employs incremental BA with a sliding window over the last 10 keyframes. Finally, in [10], a recent semi-dense approach is used to estimate the depth map of each keyframe. In this paper, we propose a different VO approach which is based on the so-called *Known Rotation Problem* [5]. It allows us to globally (i.e not incrementally) estimate submaps of keyframes while efficiently rejecting outlier tracks thanks to a Linear Program (LP).

2) A loop closure module which prevents scale drift. It consists in detecting loop closures between the submaps as in [6, 7] (or directly between the keyframes as in [9, 10]) and minimizing a cost function. To do so, [6] employs the hierarchical framework of [11], [7] uses Preconditioned Gradient Descent while [9] and [10] apply a Levenberg-Marquardt algorithm. Also, none of the previous previously cited methods deal with erroneous loop closures which increases their chances of failure especially in large environments. Contrary to those approaches, we propose a *loopy belief propagation* algorithm which is able to efficiently handle a large number of loop closures without the need of any initialization. Furthermore, we propose a simple and efficient *outlier removal algorithm* which is able to reject false loop closures.

The framework proposed in [12] is closely related to the one proposed in [9]. However several modifications have been proposed and they achieve the state of the art results on the KITTI dataset. Thus, in the rest of the paper, we compare our novel monocular VSLAM framework to the state of the art algorithms [12] and [10].

3. Proposed Monocular VSLAM Framework

The proposed monocular VSLAM framework consists in 4 modules (Keyframe Selection, Submap Reconstruction, Pairwise Similarity Estimation and Relative Similarity Averaging) arranged as illustrated in Fig.1. The first three modules are presented in this section while the last one is described in the next section.

3.1. Keyframe Selection

Selecting keyframes among all the frames of a video is necessary in order to keep a reasonable computational complexity during the monocular VSLAM process. In order to select keyframes, we apply a Lucas-Kanade tracker by detecting and tracking Harris Points of Interest (PoI) in the video frames. A frame is selected as a keyframe when the Euclidean distance between the PoI of the current frame and the PoI of the previous keyframe is greater than a given threshold (typically 5% of the image width).

This algorithm allows to efficiently select keyframes for any camera motion.

3.2. Submap Reconstruction

After having selected keyframes, we define clusters of L consecutive keyframes. Then, we apply a Structure from Motion (SfM) algorithm based on the *Known Rotation Problem* [5] to each cluster independently in order to obtain submaps. The SfM algorithm we propose is similar to the one proposed in [13] yet significantly different since it deals with temporally consecutive frames and not unordered image collections. Let us now describe this SfM algorithm.

First of all, SURF PoI [14] are extracted from all keyframes. Then the SURF descriptors are matched between pairs of keyframes in order to close loops inside each submap. The epipolar geometry is robustly estimated (five point algorithm [15] combined with a RANSAC [16] algorithm and a final BA) between pairs of images using both the SURF matches and the previously tracked Harris PoI. Since the keyframes are temporally consecutive, this is only performed for a subset of pairs of images.

After that, the relative 3D orientations extracted from the epipolar matrices are used to estimate the global 3D orientations. In order to robustly estimate the global orientations, it is actually possible to employ the relative similarity averaging algorithms (Alg.1 and Alg.2 that we will describe in the next section) since a 3D orientation is simply a 3D similarity with a scale of 1 and no translation part.

Now that the global 3D orientations have been estimated, we build tracks of PoI and employ an LP to solve the *Known Rotation Problem*, i.e to estimate the camera pose of each keyframe as well as the 3D point associated to each track. Once again, this step is made robust to erroneous tracks by employing the Linear Program[1] proposed in [5]. A BA algorithm is finally applied to refine the reconstruction.

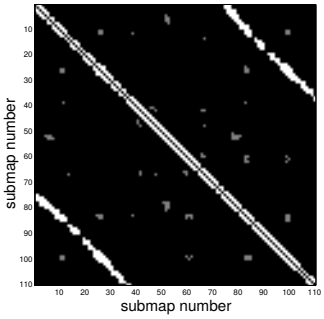
This SfM algorithm is able to robustly and accurately estimate each submap independently, even for small baselines and an environment not completely static (see section 6).

3.3. Pairwise Similarity Estimation

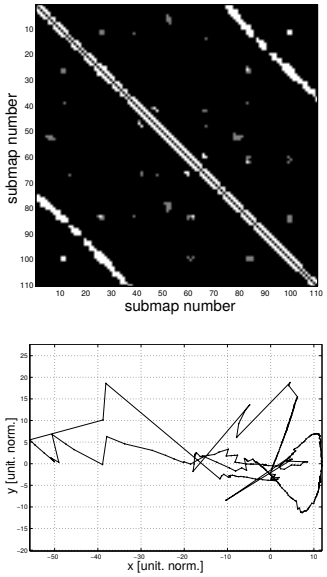
Once the submaps have been reconstructed, the loop closures are detected in two steps. First, a bag of words approach is applied to the SURF descriptors of the 3D points of all the submaps to obtain an unique descriptor for each submap. Then, a 3D similarity is estimated between each

[1] We use the MOSEK optimization toolbox for Matlab to solve the LP.

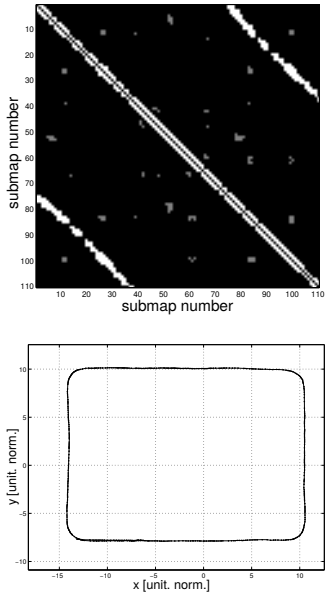
(a) Example of video frames



(b) Ground truth



(c) (top) Outlier removal result using only the temporally consecutive measurements. Some outlier are classified as inliers. (down) The trajectory is not correctly estimated.



(d) (top) Result of Alg.1. The outliers are perfectly detected, (down) The trajectory is correctly estimated (we obtain an almost perfect rectangle).

Figure 2: Example of result on a video taken in the corridor of a building (the corridor forms a rectangle) ($t_{\chi^2} = 16, n = 10$). In the labeling matrices, a white pixel is an inlier, a black pixel corresponds to an unavailable measurement and a gray pixel corresponds to an outlier.

submap and its 10 nearest neighbors. The relative 3D similarity between two submaps is estimated as follows:

1. The SURF descriptors of the 3D points of each submap are matched using a k-d tree.
2. A 3 points algorithm [17] combined with a RANSAC is applied to the matches to obtain a 3D similarity, followed by a non-linear refinement.

In all these steps, only the 3D points that have a small covariance are involved. Also, the relative similarity between two temporally consecutive submaps is always computed.

4. Large Scale Relative Similarity Averaging

After having estimated relative 3D similarities between pairs of submaps, we wish to estimate the global 3D similarities, i.e the 3D similarities between a global reference frame and the reference frame of each submap, in order to align all the submaps.

4.1. Preliminaries

4.1.1 Geometry of 3D similarities

A 3D similarity $X_{ij} = \begin{bmatrix} s_{ij}R_{ij} & T_{ij} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix} \subset \mathbb{R}^{4\times 4}$ is a transformation matrix where s_{ij} is a scale factor, R_{ij} is a 3D rotation matrix and T_{ij} is a 3D vector. Applying X_{ij} to a 3D point $x^j \in \mathbb{R}^3$ defined in a reference frame (RF) j allows to transform x^j from RF j to RF i, i.e $\begin{bmatrix} x^i \\ 1 \end{bmatrix} = X_{ij} \begin{bmatrix} x^j \\ 1 \end{bmatrix}$. Two similarities X_{ij} and X_{jk} can be composed using matrix multiplication to obtain another similarity $X_{ik} = X_{ij}X_{jk}$. Inverting a similarity matrix X_{ij} produces the inverse transformation, i.e $X_{ij}^{-1} = X_{ji}$. Consequently multiplying a transformation with its inverse produces the identity matrix: $X_{ij}X_{ji} = Id$. From a mathematical point of view, the set of 3D similarities form the 7-dimensional matrix Lie group $Sim\,(3)$ [18]. The matrix exponential exp and matrix logarithm log establish a local diffeomorphism between an open neighborhood of Id in $Sim\,(3)$ and an open neighborhood of $\mathbf{0}_{4\times 4}$ in the tan-

gent space at the identity, called the *Lie Algebra* $\mathfrak{sim}(3)$. The Lie Algebra $\mathfrak{sim}(3)$ is a 7-dimensional vector space. Hence there is a linear isomorphism between $\mathfrak{sim}(3)$ and $\mathbb{R}^7$ that we denote as follows: $[\cdot]^\vee : \mathfrak{sim}(3) \rightarrow \mathbb{R}^7$ and $[\cdot]^\wedge : \mathbb{R}^7 \rightarrow \mathfrak{sim}(3)$. We also introduce the following notations: $exp^\wedge(\cdot) = exp\left([\cdot]^\wedge\right)$ and $log^\vee(\cdot) = [log(\cdot)]^\vee$. It means that a transformation $X_{jj'}$ that is "close enough" to Id can be parametrized as follows: $X_{jj'} = exp^\wedge(\delta_{jj'}) \in Sim(3)$. Finally, we remind the adjoint representation $Ad(\cdot) \subset \mathbb{R}^{7\times7}$ of $Sim(3)$ on $\mathbb{R}^7$ that enables us to transport an increment $\epsilon_{ij}^i \in \mathbb{R}^7$, that acts onto an element X_{ij} through left multiplication, into an increment $\epsilon_{ij}^j \in \mathbb{R}^7$, that acts through right multiplication:

$$exp^\wedge\left(\epsilon_{ij}^i\right) X_{ij} = X_{ij} exp^\wedge\left(\epsilon_{ij}^j\right) \tag{1}$$

where

$$\epsilon_{ij}^j = Ad\left(X_{ij}^{-1}\right)\epsilon_{ij}^i = Ad\left(X_{ji}\right)\epsilon_{ij}^i \tag{2}$$

4.1.2 Concentrated Gaussian Distribution on $Sim(3)$

The distribution of a random variable $X_{ij} \in Sim(3)$ is called a (right) concentrated Gaussian distribution on $Sim(3)$ [19] of "mean" μ_{ij} and "covariance" P_{ij}^i if:

$$X_{ij} = exp^\wedge\left(\epsilon_{ij}^i\right)\mu_{ij} \tag{3}$$

where $\epsilon_{ij}^i \sim \mathcal{N}_{\mathbb{R}^7}\left(\mathbf{0}_{7\times1}, P_{ij}^i\right)$ and $P_{ij}^i \subset \mathbb{R}^{7\times7}$ is a definite positive matrix. Such a distribution provides a meaningful covariance representation and allows us to quantify the uncertainty of the 3D similarities.

4.2. Relative Similarity Averaging Problem

4.2.1 Without wrong loop closures

Assuming that the relative similarity measurements computed in section 3.3 do not contain wrong loop closures, the problem of relative similarity averaging consists in minimizing the following cost function:

$$\underset{\{X_{iS}\}_{i\in\mathcal{V}}}{\operatorname{argmin}} \left(\sum_{(i,j)\in\mathcal{E}} \left\| log^\vee\left(Z_{ij}X_{jS}X_{iS}^{-1}\right)\right\|_{\Sigma_{ij}^i}^2 \right) \tag{4}$$

where $\|\cdot\|_\cdot^2$ is the Mahalanobis distance, $Z_{ij} \in Sim(3)$ is a noisy relative similarity measurement between a RF j and a RF i. S is the global RF and X_{iS} and X_{jS} are the global similarities that we want to estimate. This formulation comes from the generative model:

$$Z_{ij} = exp^\wedge\left(b_{ij}^i\right) X_{iS}X_{jS}^{-1} \tag{5}$$

where $b_{ij}^i \sim \mathcal{N}_{\mathbb{R}^p}\left(\mathbf{0}_{p\times1}, \Sigma_{ij}^i\right)$ is a white Gaussian noise. In practice, the relative 3D similarity measurement Z_{ij} is obtained, as explained in section 3.3, by computing the relative 3D similarity between submaps i and j. The covariance matrix Σ_{ij}^i is obtained using a Laplace approximation after the non-linear refinement.

The problem (4) can be seen as the inference in a factor graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where each vertex $\mathcal{V}_i$ corresponds to a global similarity X_{iS} and each pairwise factor $\mathcal{E}_{ij}$ corresponds to a relative measurement Z_{ij} (see Fig.4a) which links two vertices $\mathcal{V}_i$ and $\mathcal{V}_j$.

4.2.2 In the presence of wrong loop closures

The problem (4) is based on an L2-norm and consequently is not robust to wrong loop closures. When a camera evolves in a large environment, it is common to detect wrong loop closures that leads to an outlier relative similarity. Consequently, before solving (4), we need to remove the outlier relative 3D similarity measurements in the graph.

4.3. Related Work

In [19], an Iterated Extended Kalman Filter on Lie Groups is applied to the same problem as (4). It allows to efficiently estimate both the global similarities while rejecting outliers. However, this approach cannot be applied to estimate a large number of global similarities ($N > 500$) because of the size of the covariance matrix ($7N \times 7N$).

In [20], a method based on collecting the loop errors in the graph is derived to infer the set of outliers. Nevertheless, collecting the loop errors becomes intractable and the maximum loop length is limited to 6.

Also, several recent approaches have been proposed in the field of graph-based SLAM [21, 22, 23]. These approaches employ the Levenberg-Marquardt algorithm to simultaneously perform the inference in the graph and reject outliers. However, they do not deal with 3D similarity measurements.

Contrary to these approaches, we show that by intrinsically taking into account the nature of the relative similarity averaging problem in the context of VSLAM, it is possible to separate the outlier rejection task from the inference. In the next section, we present a simple and efficient *outlier removal algorithm* while our novel *message passing algorithm* dedicated to large scale relative similarity averaging is described in section 4.5.

4.4. Outlier Removal Algorithm

In order to efficiently reject outliers, we assume that temporally consecutive measurements $Z_{(i-1)i}$ are not outliers, i.e relative similarities computed between consecutive submaps are not wrong loop closures. This is a classical assumption in robust graph based SLAM [22] which is verified in all our experiments (see section 6).

In an outlier free graph, integrating the relative similarities along a cycle results in an "small" error in the sense

Algorithm 1 Outlier Removal Algorithm

Inputs: $\{Z_{ij}\}_{1\leq i<j\leq N}$ (relative similarities), $\left\{\Sigma^i_{ij}\right\}_{1\leq i<j\leq N}$ (covariance matrices), t_{χ^2} (χ^2 p-value)

Outputs: $\mathcal{E}$ (set of inlier relative similarity measurements)

1. Initialize an empty graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$
2. Add vertex X_{1S} to $\mathcal{V}$
3. For k from 2 to N
   (a) Add X_{kS} to $\mathcal{V}$
   (b) Add $\left\{Z_{(k-1)k}, \Sigma^{k-1}_{(k-1)k}\right\}$ to $\mathcal{E}$
   (c) For each measurement Z_{lk} where $l < k$
      i. Find shortest path from X_{kS} to X_{lS} in $\mathcal{G}$
      ii. Compute the cycle error ϵ and covariance P
      iii. If $\epsilon^T P^{-1} \epsilon < t_{\chi^2}$ then add $\left\{Z_{lk}, \Sigma^l_{lk}\right\}$ to $\mathcal{E}$

that:

$$\epsilon^T P^{-1} \epsilon < t_{\chi^2} \tag{6}$$

where ϵ is the cycle error, P is its covariance and t_{χ^2} is a threshold based on the p-value of χ^2 (7) [24]. A cycle error and its covariance can be obtained efficiently using the following two equations (7) and (8) that allows to integrate/compose two relative similarities as well as their covariances.

$$Z_{kl} = Z_{km} Z_{ml} \tag{7}$$

$$\Sigma^k_{kl} \simeq \Sigma^k_{km} + Ad\left(Z_{km}\right) \Sigma^m_{ml} Ad\left(Z_{km}\right)^T \tag{8}$$

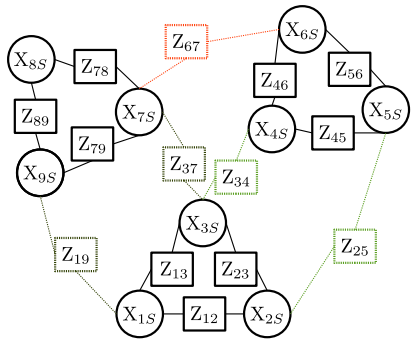
The simple method that consists in integrating the temporally consecutive measurements and checking the loop closures with (6) fails for long videos (see Fig.2c). Thus we propose an efficient algorithm (see Alg.1) that incrementally checks the loop closures by finding the shortest loop in a graph of inliers, computing the cycle error and covariance and adding the loop closure to this same graph of inliers if (6) is verified. An example of result of Alg.1 is presented in Fig.2d.

4.5. Loopy Belief Propagation

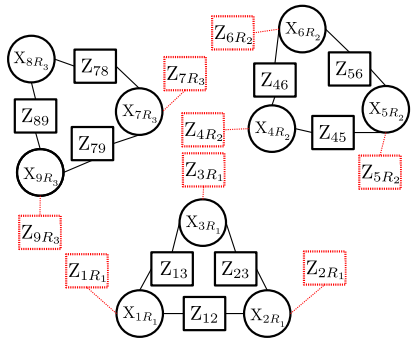
In this section, we propose a loopy belief propagation algorithm called Large Scale Relative Similarity Averaging (LS-RSA) that relies on the specific structure of the problem (4) and allows to estimate a large number of global similarities (such as $N = 10000$) very efficiently.

It consists in first partitioning the original graph $\mathcal{G}$ into N_S sub-graphs $\left\{\mathcal{G}^k = \left\{\mathcal{V}^k, \mathcal{E}^k\right\}\right\}_{k=1:N_S}$ (see Fig.4a). Then our approach alternates between:

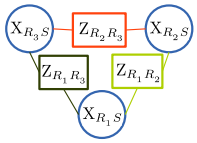
i) solving the following problem for each sub-graph $\mathcal{G}^k$ independently (see Fig.4b):



(a) Example of original factor graph. Removing the factors represented by a dotted line results in a temporally partitioned factor graph with 3 sub-graphs of size $n = 3$.



(b) Factor graph involved in step 3) of Alg.2. The subgraphs are disconnected and thus the inference can be performed in parallel.



(c) Super factor graph involved in steps 4) and 5) of Alg.2.

Figure 4: Factor graphs involved in our LS-RSA algorithm

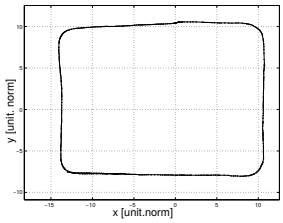
$$\underset{\left\{X_{iR_k}\right\}_{i\in\mathcal{V}^k}}{\text{argmin}} \left(\sum_{(i,j)\in\mathcal{E}^k} \left\| log^\vee\left(Z_{ij} X_{jR_k} X^{-1}_{iR_k}\right)\right\|^2_{\Sigma^i_{ij}} + \sum_{i\in\mathcal{V}^k} \left\| log^\vee\left(Z_{iR_k} X^{-1}_{iR_k}\right)\right\|^2_{\Sigma^i_{iR_k}} \right) \tag{9}$$

where each Z_{iR_k} is a global similarity measurement with covariance $\Sigma^i_{iR_k}$ that can be interpreted as a message sent from the other sub-graphs to the node $\mathcal{V}^k_i$ of $\mathcal{G}^k$;

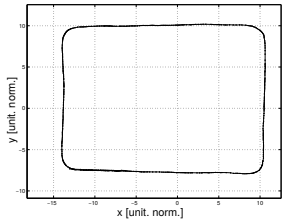
ii) computing the messages, i.e the global similarity measurements Z_{iR_k} and their covariances $\Sigma^i_{iR_k}$, by building and solving a super-graph $\mathcal{G}^{Super} = \left\{\mathcal{V}^{Super}, \mathcal{E}^{Super}\right\}$ (see Fig.4c).

Each sub-graph can be processed in parallel which makes this new message passing algorithm very efficient.

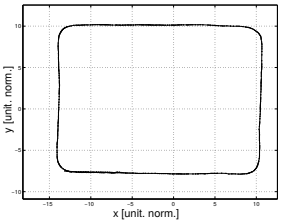
The pseudo-code of the LS-RSA is presented in Alg.2.



(a) iteration 1, residual 24.46



(b) iteration 2, residual 5.31



(c) iteration 5, residual 0.46

Figure 3: Illustrations of the iterations of the LS-RSA algorithm on the video sequence presented in Fig.2 ($n = 10$)

	Proposed	[10]	[25]	[26]	[27]	[28]
Uses Depth	No	No	No	No	Yes	Yes
fr2/desk	**2.22**	4.52	13.50	x	**1.77**	9.5
fr2/xyz	**1.28**	1.47	3.79	24.28	**1.18**	2.6

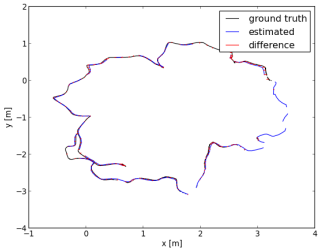


Figure 5: Left: Results on the TUM RGB-D dataset. The figures represent the absolute trajectory RMSE (cm) [29]. Right: Camera trajectory estimated with the approach proposed in this paper on the fr2/desk sequence. Notice the small error w.r.t the scale of the trajectory, hence the overlap of the curves. This plot was obtained using the online evaluation tool available on TUM RGB-D dataset webpage.

We now detail each step of this novel algorithm:

1) Graph Partitioning: The first step consists in temporally partitioning the original graph $\mathcal{G}$ into N_S sub-graphs of maximum size n where $n > 1$. In the rest of the paper, we assume, without loss of generality, that N is a multiple of N_S i.e each sub-graph has exactly n nodes. Here, the term "temporally" means that $\mathcal{G}$ is partitioned by removing the measurements that connect the following sets of nodes: $\{X_{iS}\}_{i=1:n}$, $\{X_{iS}\}_{i=n+1:2n}$, ..., $\{X_{iS}\}_{i=N-n+1:N}$ (see Fig.4a). The removed measurements are called inter-measurements.

2.a) Messages Initialization: Initialize each message Z_{iR_k} with the identity matrix and its covariance matrix $\Sigma^i_{iR_k}$ with infinite covariance and go to 3).

2.b) Messages Computation: Using the previously estimated super global similarities, we can compute the messages that are going be passed between the sub-graphs (actually between nodes of sub-graphs). A node sends a message to another if both nodes are connected by an inter-measurement. For each inter-measurement Z_{ij}, a message is created and consists in computing the global similarity measurement Z_{iR_k} and its covariance $\Sigma^i_{iR_k}$ as follows:

$$Z_{iR_k} = Z_{ij} X_{jR_l} X_{R_l S} X^{-1}_{R_k S} \tag{10}$$

$$\begin{aligned}\Sigma^i_{iR_k} = Ad_G\left(Z_{ij}\right)\Big[P^j_{jR_l} + Ad_G\left(X_{jR_l}\right)\Big\{P^{R_l}_{R_l S} \\ + Ad_G\left(X_{R_l S} X^{-1}_{R_k S}\right) P^{R_k}_{R_k S} Ad_G\left(X_{R_l S} X^{-1}_{R_k S}\right)^T\Big\} \\ Ad_G\left(X_{jR_l}\right)^T\Big] Ad_G\left(Z_{ij}\right)^T + \Sigma^i_{ij}\end{aligned} \tag{11}$$

This can be interpreted as a message sent from X_{jR_l} to X_{iR_k}. If a node X_{iR_k} receives multiple messages i.e several Z_{iR_k} have been computed because X_{iR_k} is connected to several inter-measurements, we apply a Karcher mean (see [30] section IV) to summarize those messages into a single one.

3) Subgraphs Optimization: For each sub-graph $\mathcal{G}^k$, we estimate the global similarities $\{X_{iR_k}\}_{i\in\mathcal{V}^k}$ as well as the marginal covariances of the posterior distribution $\left\{P^i_{iR_k}\right\}_{i\in\mathcal{V}^k}$ by applying a Levenberg-Marquardt to solve (9) followed by a Laplace approximation. Note that if there is only one sub-graph ($N_S = 1$), then the algorithm stops and returns the result. This step is illustrated by Fig.4b.

4) Super Graph Building: We now build a super graph $\mathcal{G}^{Super} = \left\{\mathcal{V}^{Super}, \mathcal{E}^{Super}\right\}$ (see Fig.4c) from the output of step 3) and the inter-measurements. The edges of this super graph are relative similarities between the reference frames $\{R_k\}_{k=1:N_S}$ called super-measurements. Each inter-measurement Z_{ij} with covariance Σ^i_{ij} leads to the following super-measurement:

$$Z_{R_k R_l} = X^{-1}_{iR_k} Z_{ij} X_{jR_l} \tag{12}$$

with covariance matrix:

$$\begin{aligned}\Sigma^{R_k}_{R_k R_l} = Ad_G\left(X^{-1}_{iR_k}\right)\left(P^i_{iR_k} + \Sigma^i_{ij}\right. \\ \left. + Ad_G\left(Z_{ij}\right) P^j_{jR_l} Ad_G\left(Z_{ij}\right)^T\right) Ad_G\left(X^{-1}_{iR_k}\right)^T\end{aligned} \tag{13}$$

Since each inter-measurement leads to a super-measurement, we may have several super-measurements

Algorithm 2 Large Scale Relative Similarity Averaging (LS-RSA)

Inputs: $\{Z_{ij}\}_{1\leq i<j\leq N}$ (relative similarities), $\{\Sigma^i_{ij}\}_{1\leq i<j\leq N}$(covariance matrices), n (subgraph size)

Outputs: $\{X_{iS}\}_{1\leq i\leq N}$ (global similarities), $\{P^i_{iS}\}_{1\leq i\leq N}$ (marginal covariance matrices of global similarities)

1) Partition the graph.
2.a) Initialize the messages and go to 3).
2.b) Compute the new messages.
3) Solve eq.(9) for each subgraph.
4) Build the super graph.
5) Apply LS-RSA (recursive call) to the super graph.
6) Compute the quantities of interest, i.e $\{X_{iS}\}_{i\in\mathcal{V}}$ and $\{P^i_{iS}\}_{i\in\mathcal{V}}$, and the residual of eq.(4). If the residual has not been reduced, then return, else go to 2.b).

between two nodes of the super graph. When it happens, we average these super-measurements using a Karcher mean to get only one super-measurement (and its covariance matrix) between two nodes.

5) Super Graph Optimization: Once the super-graph is built, we apply the LS-RSA algorithm in order to obtain the super global similarities $\{X_{R_kS}\}_{k=1:N_S}$ and $\left\{P^{R_k}_{R_kS}\right\}_{k=1:N_S}$. Note that during this step, the LS-RSA algorithm is recursively called until step 1) produces only one sub-graph (in this case, the LS-RSA exits at step 3)).

6) The quantities of interest, i.e $\{X_{iS}\}_{i\in\mathcal{V}}$ and $\{P^i_{iS}\}_{i\in\mathcal{V}}$ can be obtained as follows:

$$X_{iS} = X_{iR_k}X_{R_kS} \tag{14}$$

$$P^i_{iS} = P^i_{iR_k} + Ad_G\left(X_{iR_k}\right)P^{R_k}_{R_kS}Ad_G\left(X_{iR_k}\right)^T \tag{15}$$

and the residual of (4) can be computed. If this residual is higher than the one computed at the previous iteration, i.e the error has been not reduced, then the algorithm exits. Otherwise go to 2.b).

In section 6, we demonstrate on a large number of video sequences that the LS-RSA algorithm provides an accurate solution to the relative similarity averaging problem. A thorough analysis of the convergence properties of the proposed algorithm is beyond the scope of this paper and remains for future work.

5. Limitations

The monocular VSLAM framework proposed in this paper has several limitations. First of all, in each submap, the camera motion must not be a pure rotation otherwise the 3D point cloud cannot be estimated. However, due to the use of the known rotation problem, the proposed framework can provide accurate estimates even for small translations of the camera. Secondly, the 3D environment to be reconstructed should be static. Nevertheless, the robustness of our framework allows some objects to move in the environment, such as cars in the KITTI dataset (see section 6.2). Thirdly, if the temporally consecutive relative similarities are not outlier free, then Alg.1 will not remove all the outliers. However, in all our experiments it never happened. Finally, our monocular VSLAM approach is currently coded in Matlab (which will be made publicly available) and the processing time is 2.5 hours for a video sequence of 10000 frames.

6. Experiments

In this section, we compare the performances of the proposed approach to the state of the art algorithms [10] and [12] on the TUM and KITTI datasets as well as on several other challenging videos. For all these experiments, the parameters of our novel approach have been optimized by hand on one video once and for all ($L = 16$, $n = 10$, $t_{\chi^2} = 16$, as well as several other parameters such as RANSAC thresholds). Increasing L over 16 did not led to notable improvements. Also, in monocular VSLAM the 3D model and the camera trajectory are estimated up to a scale factor. Thus, in the rest of this section, when evaluating the results of the different approaches w.r.t the ground truth, we estimate a 3D similarity by minimizing the distance between the estimated camera trajectory and the ground truth camera trajectory.

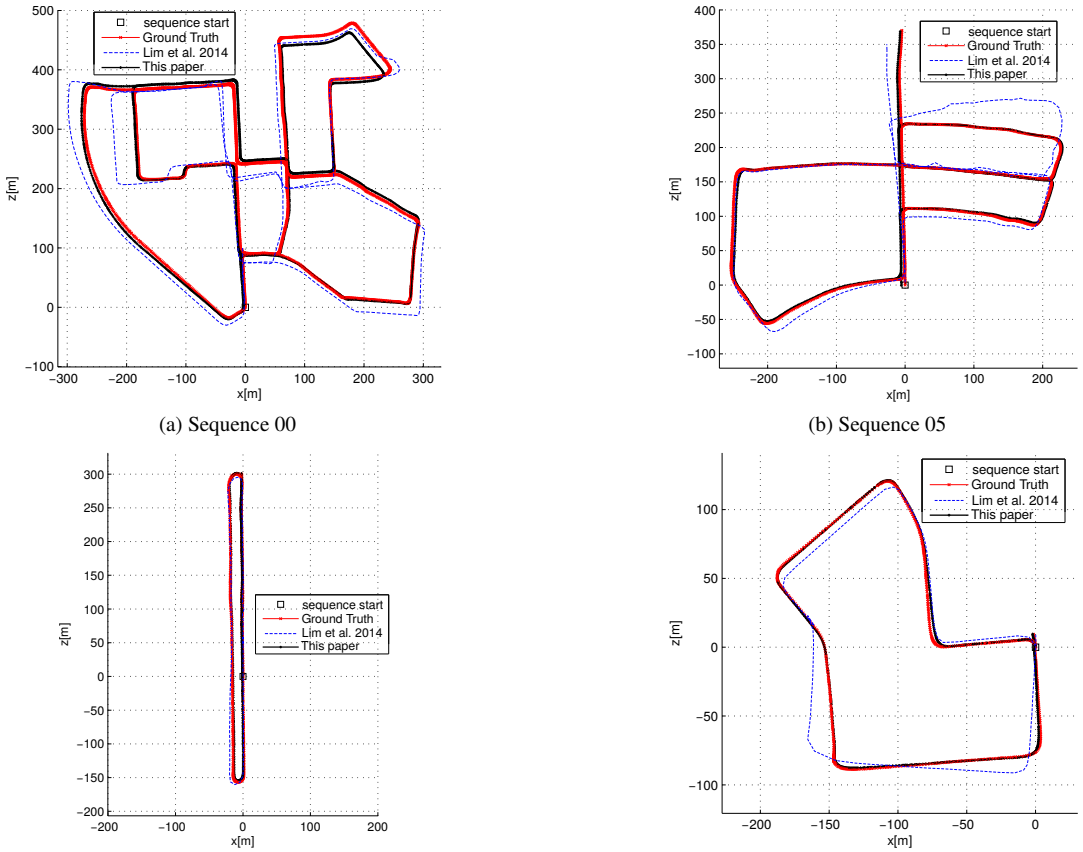
6.1. Quantitative comparison on the TUM dataset

In [10], the TUM RGB-D dataset [29] is used to evaluate their algorithm. Thus, we chose this same benchmark to quantitatively evaluate the performances of our approach w.r.t [10]. In Fig.5 (left), the absolute trajectory RMSE (cm) [29] of our novel framework, [10], [25], [26], [27] and [28] are presented[2]. For each video sequence, our approach produces a lower RMSE than the state of the art algorithm [10]. The superior performances of our approach are probably due to the fact that in our approach a submap is estimated from several keyframes while in [10], once a keyframe is selected, the semi-dense depth map of the previous keyframe is not updated anymore. It can also be seen that the results of our approach tend towards the results of the state of the art RGB-D SLAM algorithm [27] that uses an RGB-D camera instead of a classical monocular RGB camera. In Fig.5 (right), the camera trajectory estimated with our approach on the FR2/desk sequence is presented.

6.2. Qualitative comparison on the KITTI dataset

In [12], the KITTI dataset [31] is used to evaluate their algorithm. Thus, we chose this same benchmark to qualita-

[2]The results of [10, 25, 26, 27, 28] are taken from Fig.9 in [10].



(a) Sequence 00

(b) Sequence 05

(c) Sequence 06

(d) Sequence 07

Figure 6: Qualitative comparison on the camera trajectories estimated with the approach proposed in this paper and [12] on several sequences of the KITTI dataset. Most of the time, the camera trajectory estimated with our approach overlaps with the ground truth as opposed to [12] which deviates from the real trajectory.

tively evaluate the performances of our approach w.r.t [12]. In Fig.6 the camera trajectories estimated with our approach and with [12] are compared to the ground truth trajectories. On each of these plots the camera trajectory estimated by our framework is closer to the ground truth than the one estimated by [12]. This is probably due to the fact that in all these video sequences, the environment is not completely static (cars are moving). Consequently, our framework which has been tailored to be robust outperforms [12].

6.3. Qualitative comparison on challenging videos

Let us now present the results of our approach on challenging videos taken from a rolling shutter camera. The videos are corrupted by motion blur, the environment is sometimes poorly textured and the camera trajectories contain small camera translations. In Fig.2d, we show the estimated camera trajectory along the corridor of a building (the corridor forms a rectangle). One can see that the estimated trajectory is almost perfectly rectangular and flat. On that video sequence the semi-dense tracker of [10] fails. Due to the lack of space, results on other video sequences are provided as supplementary material.

7. Conclusion

The contribution of this paper is threefold:

1. A novel visual odometry approach based on the so-called *Known Rotation Problem* that allows to robustly estimate each submap independently.
2. A simple and efficient *outlier removal algorithm* to reject the outlier relative 3D similarities coming from wrong loop closures.
3. A *loopy belief propagation* algorithm which is able to align a large number of submaps very efficiently.

Using state of the art tools coming from the field of SfM from unordered image collections, we proposed a novel robust monocular VSLAM framework which is able to operate on long challenging videos. The method has been validated experimentally and compared to the two most recent state of the art algorithms which it outperforms both qualitatively and quantitatively. Moreover, in all our experiments (4 different cameras with different resolutions), the parameters of our method have been set once and for all proving the flexibility of the proposed approach.

Acknowledgments The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007- 2013) under grant agreement 288199 - Dem@Care. The authors would like to thank the reviewers and Cornelia Vacar for their valuable help as well as Carl Olsson and Jakob Engel for making their code available.

References

[1] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2d-to-3d matching," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 667–674. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6126302 1

[2] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 15–29. 1

[3] K. Konolige and M. Agrawal, "Frameslam: From bundle adjustment to real-time visual mapping," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1066–1077, 2008. 1

[4] C. Mei, G. Sibley, and M. Cummins, "A constant-time efficient stereo slam system." in *BMVC 2009*. 1

[5] C. Olsson, A. Eriksson, and R. Hartley, "Outlier removal using duality," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1450–1457. 1, 2, 3.2

[6] L. A. Clemente, A. J. Davison, I. D. Reid, J. Neira, and J. D. Tardós, "Mapping large loops with a single hand-held camera." in *Robotics: Science and Systems*, vol. 2, 2007. 2

[7] E. Eade, "Monocular simultaneous localisation and mapping," Ph.D. dissertation, 2008. 2

[8] E. Eade and T. Drummond, "Monocular SLAM as a graph of coalesced observations," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8. 2

[9] H. Strasdat, J. Montiel, and A. Davison, "Scale drift-aware large scale monocular SLAM," in *RSS*, 2010. 2

[10] J. Engel, T. Schops, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," *ECCV, Lecture Notes in Computer Science*, pp. 834–849, 2014. 2, **??**, 6, 6.1, 2, 6.3

[11] C. Estrada, J. Neira, and J. D. Tardós, "Hierarchical SLAM: real-time accurate mapping of large environments," *Robotics, IEEE Transactions on*, vol. 21, no. 4, pp. 588–596, 2005. 2

[12] H. Lim, J. Lim, and H. J. Kim, "Real-time 6-dof monocular visual SLAM in a large-scale environment," in *ICRA*, 2014. 2, 6, 6.2, 6

[13] C. Olsson and O. Enqvist, "Stable structure from motion for unordered image collections," in *Image Analysis*. Springer, 2011, pp. 524–535. 3.2

[14] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 404–417. 3.2

[15] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 6, pp. 756–777, 2004. 3.2

[16] 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. 3.2

[17] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 13, no. 4, pp. 376–380, 1991. 2

[18] G. S. Chirikjian, *Stochastic Models, Information Theory, and Lie Groups, Volume 2*. Springer-Verlag, 2012. 4.1.1

[19] G. Bourmaud, R. Mégret, A. Giremus, and Y. Berthoumieu, "Global motion estimation from relative measurements in the presence of outliers," ACCV 2014. 4.1.2, 4.3

[20] C. Zach, M. Klopschitz, and M. Pollefeys, "Disambiguating visual relations using loop constraints," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1426–1433. 4.3

[21] N. Sunderhauf and P.Protzel, "Switchable constraints for robust pose graph SLAM," in *IROS*, 2012. 4.3

[22] Y. Latif, C. Cadena, and J. Neira, "Robust loop closing over time for pose graph SLAM," *The International Journal of Robotics Research*, 2013. 4.3, 4.4

[23] E. Olson and P. Agarwal, "Inference on networks of mixtures for robust robot mapping," in *RSS*, 2012. 4.3

[24] R. A. Fisher, F. Yates *et al.*, "Statistical tables for biological, agricultural and medical research." *Statistical tables for biological, agricultural and medical research.*, no. Ed. 3., 1949. 4.4

[25] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1449–1456. **??**, 6.1, 2

[26] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *International Symposium on Mixed and Augmented Reality (ISMAR)*, 2007. **??**, 6.1, 2

[27] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 2100–2106. **??**, 6.1, 2

[28] F. Endres, J. Hess, N. Engelhard, J. Sturm, D. Cremers, and W. Burgard, "An evaluation of the RGB-D SLAM system," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1691–1696. **??**, 6.1, 2

[29] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 573–580. 5, 6.1

[30] T. D. Barfoot and P. T. Furgale, "Associating uncertainty with three-dimensional poses for use in estimation problems," *IEEE Trans. Robot.*, vol. 30, no. 3, pp. 679–693, Jun 2014. 4.5

[31] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013. 6.2