Robustness in View-Graph SLAM

Tariq Abuhashim and Lorenzo Natale
iCub Facility
Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genova, Italy.
{tariq.abuhashim, lorenzo.natale}@iit.it

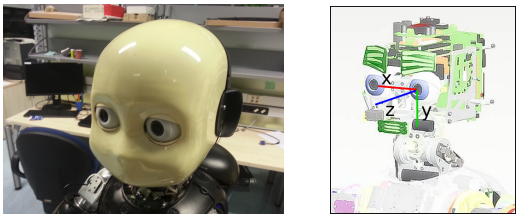
***Abstract*—This paper presents a new robustification procedure for nonlinear least-squares optimisation problems. In particular, we focus on the robustness of view-graph SLAM against outlier correspondences in the images and outlier geometries in the graph. Our method utilises revised measurements model linearisation and decision making to detect and remove outliers during data fusion. We utilise innovations and residuals gating to decide which observations were affected, given the most recent model linearisation point. By exploiting the inherited locality of measurements and states and the sparsity structure of nonlinear least-squares formulation we can check which measurement is affected before and after data fusion. This locality is the whole basis of robustification. To be efficient, we carry out the estimation using the information form. By doing so, we can include and remove information by individual measurements at each step using simple information addition and subtraction operations. Our results demonstrate the robustness of our method against outliers with respect to the use of kinematics alone and RANSAC with Levenberg-Marquardt algorithm.**

I. Introduction

Using an extrinsically calibrated stereo pair is a common solution to obtain reliable localisation and mapping results (see for instance [1]). In order to obtain accurate depth estimates, the cameras are usually separated by a significant baseline thus creating widely spaced observations of the same object. In certain practical cases, however, the distance between the cameras is small. In these cases the problem become difficult because small correspondence error can affect the depth estimation significantly.

In this paper, we address the problem of visual localisation and mapping using iCub (shown in figure 1). A humanoid robot which mounts two cameras separated by a small baseline[1]. The two cameras are mounted on a platform where their rotations around the x and z axes are coupled, while they may rotate independently around the y axis. Even though an estimate of the relative motion between the two cameras can be computed using a kinematic model, this estimate is unreliable due to model errors, outliers, and triangulation uncertainty. Hence, we use nonlinear least-squares to estimate the motion parameters of each camera from the images, while kinematics are used to initialise the map and hence to solve for the scale ambiguity. This generalises our method to various visual systems, where an initial scale can be obtained for instance using inertial sensors or platform kinematics.

[1]The displacement between the two cameras along the x-axis is, approximately, 68 mm. Which is small when compared to the observed distances.



(a) iCub. (b) Reference frame.

Fig. 1. iCub mounts two Dragonfly-2 color video cameras (resolution 640×480) and a PC104 computer capturing image frames at 30Hz.

Many estimation problems in robotics rely on solving nonlinear least-squares. For Simultaneous Localisation and Mapping (SLAM), or many other mathematically related formulations including Bundle-Adjustment (BA) and Structure from Motion (SfM), nonlinear least-squares solvers are getting more and more efficient [2], [3], [4]. Nevertheless, dealing with outliers due to wrong data associations and degenerate camera configurations is still an issue, and may result into delayed convergence and loss in accuracy. Hence, this paper presents an efficient and robust nonlinear least-squares estimation framework for view-graph SLAM. Our goal is to demonstrate how each of its components can be made robust to outliers in the data.

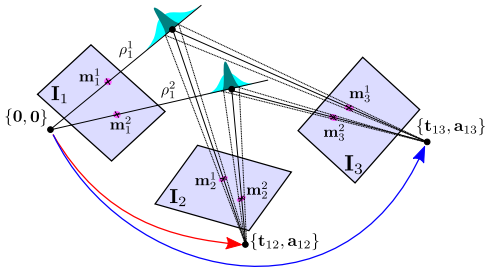
II. Problem Formulation

For view-graph SLAM the estimated state contains a set of camera poses with no map. Instead, the image frames are used to break up the map into a number of visual scans, where every scan relates to a reference camera frame. Therefore, the estimated state is the pose of the reference frame at each moment a scan is obtained. The set of visual scans can later generate a map, by projecting each one of them into the common global coordinate frame according to its pose estimate.

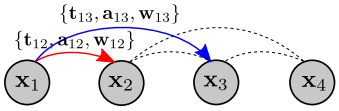
For monocular and uncalibrated stereo systems, the graph constraints and the visual scans have to be estimated before graph optimisation takes place. This is *a two-step procedure*, as shown in figure 2.

A. Pair-wise States and Measurements

The aim of the first estimation task is to use information from images to estimate the relative pose parameters between pairs of images. Given a set of M image frames $\{\mathbf{I}_1, \ldots, \mathbf{I}_M\}$, with N 2D correspondences $\{\mathbf{m}^1_{1:N}, \ldots, \mathbf{m}^M_{1:N}\}$, our state



(a) Step 1: Relative pose estimation.



(b) Step 2: Global pose estimation.

Fig. 2. Two step procedure for view-graph SLAM, using a graph with three-view constraints. Top: estimation of relative poses between frames. Bottom: global pose estimation, in which each frame is projected into a common global frame. Notice that for global pose estimation redundant information exists (e.g. given a reference $\mathbf{x}_1$, the state $\mathbf{x}_3$ can be obtained directly from $\mathbf{x}_{13} = \{\mathbf{t}_{13}, \mathbf{a}_{13}\}$ or by concatenating $\mathbf{x}_{12}$ and $\mathbf{x}_{23}$).

vector contains the relative translation and rotation vectors $\{\mathbf{t}_{1j}, \mathbf{a}_{1j}\}$ and N inverse depth parameters $\rho^1_{1:N} = \{\rho^1_1, \ldots, \rho^1_N\}$, computed with the frame $\mathbf{I}_1$ as a reference.

$$\mathbf{x} = \left[\mathbf{0}_{1\times3}, \mathbf{0}_{1\times3}, \mathbf{t}_{12}, \mathbf{a}_{12}, \ldots, \mathbf{t}_{1M}, \mathbf{a}_{1M}, \rho^1_{1:N}\right]^\top. \quad (1)$$

Where $\mathbf{a}_{1j}$ is parametrised following Rodrigues notation[2]. Thus, the dimensions of our state vector are $(6M + N)\times1$.

Our observations, then, consist of a set of 2D pixel locations in *normalised image coordinates*[3], which were extracted in the reference image $\mathbf{I}_1$, and then located in the remaining $M-1$ images,

$$\mathbf{z} = \left[\mathbf{m}^1_{1:N}, \ldots, \mathbf{m}^M_{1:N}\right]^\top. \quad (2)$$

The dimensions of the observations vector depends on the visibility of the map features in the images. If all the features were tracked successfully in all the M image, they should be $(MN\times1)$. These observations will form the constraints in our batch nonlinear least-squares estimation problem.

We assume a nonlinear image formation model,

$$\mathbf{h}(\mathbf{x}) \approx \mathbf{C}_1^{j\top}\left(\mathbf{x}^1_\mathbf{f} - \mathbf{t}_{1j}\right). \quad (3)$$

where $\mathbf{C}_1^j$ is a rotation matrix computed from $\mathbf{a}_{1j}$,

$$\mathbf{C}_1^j = \mathbf{I} + [\mathbf{n}]_\times \sin\|\mathbf{a}_{1j}\| + [\mathbf{n}]^2_\times\left(1-\cos\|\mathbf{a}_{1j}\|\right), \quad \mathbf{n} = \frac{\mathbf{a}_{1j}}{\|\mathbf{a}_{1j}\|},$$

[4]and $\mathbf{x}^1_\mathbf{f}$ represents 3D corners computed from inverse depth states $\rho^1_{1:N}$,

$$\mathbf{x}^1_\mathbf{f} = \frac{1}{\rho^1_n\sqrt{x_n^2 + y_n^2 + 1}}\left[x_n, y_n, 1\right]^\top, \mathbf{m}^1_n = \left[x_n, y_n\right]^\top.$$

[2]A downside to a quaternion parameterisation within a weighted-least squares state vector is that the unit norm cannot be guaranteed when calculating the mean. Thus, generalised Rodrigues parameters that represent the local rotation is preferred when constructing the data fusion equations (The details are omitted for brevity). Also, this is faster and more accurate than using Euler angular parameterisation because trigonometric functions are avoided.

[3]obtained after applying each camera intrinsic parameters.

[4]$[\cdot]_\times$ computes a skewed-symmetric matrix.

Due to errors in initial pose and inverse depth states and the noise in the images, the observed pixel locations $\mathbf{z}$ will not match the predicted locations by the nonlinear image reprojection model. Notice that, by estimating all the relatives in batch, we are technically performing local bundle-adjustment. Finally, the estimated relatives are then inserted into their corresponding edges in the view-graph, as shown in figure 2.

B. View-Graph States and Measurements

The aim of the second estimation task is to solve for the global poses of the image frames in the view-graph, given all the estimated relative poses in the previous section. Notice that, inverse depth parameters are not considered during graph estimation. However, they will be used later to build the map by projecting each inverse depth estimate into the common global frame given the estimated pose of its reference frame.

The state vector of the view-graph is composed of K camera frames,

$$\mathbf{x} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_K\right]^\top, \quad \mathbf{x}_k = \left[\mathbf{t}_k, \mathbf{a}_k\right], \quad k = 1, \ldots, K. \quad (4)$$

where $\{\mathbf{t}_k, \mathbf{a}_k\}$ are the position and orientation (parametrised using Rodrigues rotations) of the k^{th} camera frame relative to the global reference. Thus, the dimensions of our state vector are $(6K \times 1)$.

For any two camera frames in the graph, $\mathbf{x}_i$ and $\mathbf{x}_j$, the relative pose measurements and constraint model are given, respectively, by

$$\mathbf{z}_{ij} = \left[\mathbf{t}_{ij}, \mathbf{a}_{ij}\right]^\top, \quad \mathbf{h}(\mathbf{x}) \approx \left[\tilde{\mathbf{t}}_{ij}, \tilde{\mathbf{a}}_{ij}\right]^\top, \quad (5)$$

where $\tilde{\mathbf{t}}_{ij}$ is the predicted relative translation between $\mathbf{t}_i$ and $\mathbf{t}_j$, given by,

$$\tilde{\mathbf{t}}_{ij} = \mathbf{C}_w^{i\top}(\mathbf{t}_j - \mathbf{t}_i).$$

While $\tilde{\mathbf{a}}_{ij}$ is the predicted relative rotation between $\mathbf{a}_i$ and $\mathbf{a}_j$, which can be robustly computed using the corresponding rotation matrices $\mathbf{C}^i_w$ and $\mathbf{C}^j_w$,

$$\tilde{\mathbf{a}}_{ij} = \Psi(\tilde{\mathbf{C}}^j_i), \quad \tilde{\mathbf{C}}^j_i = \mathbf{C}_w^{i\,-1}\mathbf{C}^j_w.$$

Here, w denotes a global reference frame, and $\Psi(\cdot)$ performs a parameteric transformation from rotation matrix space to Rodrigues rotation space. The complete graph measurements vector is then composed of all the observed relative poses,

$$\mathbf{z} = \left[\mathbf{z}_{i,j}, \ldots, \mathbf{z}_{K-1,K}\right]^\top. \quad (6)$$

III. Related Work on The Robustification of Visual Mapping

A. Dealing with outlier image correspondences

When solving for the relative poses, tracked image points will act as measurements. Errors in relative pose and inverse depth computations are due to outlier point tracks and poor pose initialisation. This is because accurate inverse depth and relative pose computations require quality inlier matches, while finding an acceptable set of inlier matches require good relative poses. This problem is usually solved using

RANSAC iterations and epipolar geometry by the essential matrix. During every iteration, relative poses are initialised from the epipolar geometry of a randomly selected set of five matches [5]. The geometry that generates the largest number of positive inverse depth values is selected [6]. Finally, the relative pose and inverse depth parameters are refined using bundle-adjustment [7].

It is mostly due to small translations, however, that decisions made from some geometries become unreliable. This is because the accuracy of depth computation is inversely proportional to the relative frame translation and is inversely quadratic with the observed distance. In similar situations, there is no way to determine how many iterations are required in order to select an acceptable initial set of inliers with a spatial distribution that is good enough for accurate, repeatable and robust essential matrix computations. Hence, several thousands of RANSAC iterations are usually used. Also, bundle-adjustment (using Levenberg-Marquardt for instance) requires a close initialisation and, in case of poor initialisation, it is likely to remain trapped in the closest minima. Finally, inverse depth and relative translation parameters can be computed from the essential matrix only up to a scale factor. This scale ambiguity increases the complexity of outliers detection and removal.

B. Dealing with outlier graph constraints

When solving for the global poses, the previously estimated relative motions *and their corresponding uncertainty* in the view-graph will act as constraints. This is often referred to as *trajectory smoothing* [4] or *motion averaging* [8]. They are often used as pre-conditioners or initialisers to global optimisation methods, including bundle-adjustment [9].

Some of pair-wise geometries may not be solvable, for instance, small translations or degenerate configurations. During averaging, outliers in the graph will have a negative influence by gravitating the state estimate towards them. Hence, a few methods have been developed to incorporate robustness into averaging of graph constraints. The common approach is to decouple rotations from translations, and hence solve for global rotations robustly first [11].

To incorporate robustness in the averaging of rotations, available methods can be classified into two approaches. The first approach is to detect outliers in the set of relative rotations and remove them before carrying out averaging. The common approach is to use RANSAC to detect and remove outlier rotations in the graph [10], [11]. This is based on the fact that, if there is no noise or outliers, a loop of rotations in the graph should result in the identity transformation. However, RANSAC-based methods suffer from an increased computational complexity with the increased size of the view-graph. One other problem with using RANSAC is that old decisions, once made, are not revised. Also, RANSAC fails to account for measurements and state uncertainty.

The second approach is to robustly solve for global rotations without the need to explicitly detect or remove outliers. For instance, DISCO [13] treats the problem as a discrete labeling instance using Markov Random Fields, which is expected to be extremely expensive and will require a significant amount of memory. Alternatively, [14] performs robust global averaging of rotations using their corresponding Lie-algebra. This approach uses two optimisation steps, where an l_1 optimiser is used to initialise an estimate. This estimate is then refined, using all the constraints, by an iteratively re-weighted least squares (IRLS) optimiser using an l_2 cost function with a Huber-like loss function. This allows for fine tuning of the estimate while under-weighting inconsistent measurements. IRLS is a greedy algorithm and needs a good initial guess. Without a good initial guess, the intermediate weighting of the loss function will not be informative and the algorithm may not converge to a good final estimate. Also, the method was only applied for rotations, and does not consider uncertainty.

For translations, 1DSfM [15] solves first for global rotations using the method in [14], then for global translations, by optimising an objective function that depends only on comparing measurement directions to model directions. Information about which measurements are likely inconsistent is recovered by solving for multiple 1D ordering problems. This can be done as an instance of Minimum Feedback Arc Set (MFAS), a well know problem in graph theory [16].

C. Our Contributions

This paper presents a robust framework for view-graph SLAM. By doing so, we pay close attention to wrong correspondence information between the images, and then to outlier motion estimates in the whole graph. Our method utilises sparse representation, revised linearisation, and statistical testing using innovations and residuals analysis to detect outliers in the data and hence remove them. Thus, robustness against poor initial decisions about outlier measurements is acheived during data fusion by using iterative measurements model linearisation and measurements switching.

In comparison to related approaches, our approach has the following advantages:

- Unlike RANSAC-based methods in [5] and [11], our method utilises measurements and states uncertainty and allows for revised decisions once more information become available.
- Unlike the methods in [14] and [15], we couple decisions on outlier rotations and translations. This is motivated by the fact that an error in relative pose computations makes the whole geometry unreliable.
- Our robustification approach can be generalised to various nonlinear least-squares problems in robotics and computer vision.

IV. Robust Nonlinear Least-Squares Estimation

A. Optimally Weighted Non-linear Least-Squares

We are concerned with applying Gaussian estimators to systems with nonlinear models of the form,

$$\hat{\mathbf{z}} = \mathbf{h}(\mathbf{x}) + \mathbf{n}.$$

This models the predicted measurements $\hat{\mathbf{z}}$ made given the state $\mathbf{x}$, where $\mathbf{n}$ is an additive zero-mean Gaussian noise

with covariance $\mathbf{R}$. The nonlinear model can be approximated by a tangent about a linearisation point $\mathbf{x}^s$, and thus is approximated by a set of linear models,

$$\hat{\mathbf{z}} \triangleq \mathbf{h}(\mathbf{x}^s) + \nabla \mathbf{h}_{\mathbf{x}^s}(\hat{\mathbf{x}} - \mathbf{x}^s),$$

where $\nabla \mathbf{h}_{\mathbf{x}^s}$ is the Jacobian of $\mathbf{h}(\mathbf{x})$ evaluated at an arbitrarily chosen linearisation point $\mathbf{x}^s$.

The aim of smoothing and mapping algorithms is, using the observation vector $\mathbf{z}$, to estimate the state $\hat{\mathbf{x}}$ which minimises the weighted Sum of Squared Error (SSE) cost function:

$$\mathbf{f}(\mathbf{x}) = \frac{1}{2}(\mathbf{z} - \mathbf{h}(\mathbf{x})^\top \mathbf{R}^{-1}(\mathbf{z} - \mathbf{h}(\mathbf{x})),$$

where $\mathbf{R}^{-1}$ is the optimal weight matrix. This type of problem formulation (generally referred to as non-linear least squares) is typically solved using numeric optimisation methods such as Newtons method, the Gauss-Newton approximation and the Levenberg-Marquardt algorithm (which is commonly used in vision-based bundle-adjustment) [7]. The Gauss-Newton approximation of this nonlinear least-squares problem has the following solution,

$$\hat{\mathbf{x}} = \mathbf{x} + \Delta \mathbf{x}, \tag{7}$$

where,

$$\Delta \mathbf{x} = \mathbf{Y}_{\mathbf{z}}^{-1} \mathbf{y}_{\mathbf{z}},$$

$$\mathbf{Y}_{\mathbf{z}} = \nabla \mathbf{h}_{\mathbf{x}^s}^\top \mathbf{R}^{-1} \nabla \mathbf{h}_{\mathbf{x}^s}, \quad \mathbf{y}_{\mathbf{z}} = \nabla \mathbf{h}_{\mathbf{x}^s}^\top \mathbf{R}^{-1} \nu, \tag{8}$$

and ν are the innovations, $\nu \triangleq \mathbf{z} - \hat{\mathbf{z}}$. Here, $\mathbf{y}_{\mathbf{z}}$ and $\mathbf{Y}_{\mathbf{z}}$ are the information vector and matrix by the measurements $\mathbf{z}$.

By assuming that $\mathbf{x}$ is Gaussian with covariance matrix $\mathbf{P}$, we can also compute the uncertainty in the least-squares estimate $\hat{\mathbf{x}}$,

$$\hat{\mathbf{P}} = \mathbf{P} - \mathbf{P} \nabla \mathbf{h}_{\mathbf{x}^s}^\top \mathbf{S}_\nu^{-1} \nabla \mathbf{h}_{\mathbf{x}^s} \mathbf{P}, \tag{9}$$

where $\mathbf{S}_\nu$ is the innovations covariance matrix, given as,

$$\mathbf{S}_\nu = \nabla \mathbf{h}_{\mathbf{x}^s} \mathbf{P} \nabla \mathbf{h}_{\mathbf{x}^s}^\top + \mathbf{R}.$$

This propagates the uncertainty in the prior estimate $\mathbf{x}$, through the nonlinear model $\mathbf{h}(\mathbf{x})$ using a set of linearised Gaussian projections by the Jacobian $\nabla \mathbf{h}_{\mathbf{x}^s}$, which is computed at a linearisation point $\mathbf{x}^s$. From equations (7), (8), and (9), we may also write the update equations in the following information form (or inverse-covariance form)[5],

$$\hat{\mathbf{y}} = \mathbf{y} + \mathbf{y}_{\mathbf{z}},$$

$$\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{Y}_{\mathbf{z}},$$

where $\mathbf{y}$ and $\mathbf{Y}$ are the prior information vector and matrix computed using,

$$\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}, \quad \mathbf{Y} = \mathbf{P}^{-1}.$$

The least-square state estimate $\{\hat{\mathbf{x}}, \hat{\mathbf{P}}\}$ can be then recovered from $\{\hat{\mathbf{y}}, \hat{\mathbf{Y}}\}$ by solving the following sparse system,

$$\hat{\mathbf{Y}}\hat{\mathbf{x}} = \hat{\mathbf{y}}, \quad \hat{\mathbf{P}} = \hat{\mathbf{Y}}^{-1}.$$

[5]The details of this derivation are omitted for brevity.

Algorithm 1 summarises steps for solving nonlinear least-squares problems using the information form equations. It is easy to say that data fusion equations in the information form are simpler than those in the covariance form, since simple additions are employed instead of multiplications. They are also more efficient since the information matrix is typically sparse and positive definite.

1- Initialise information vector and matrix $\{\mathbf{y}, \mathbf{Y}\}$.
2- Set linearisation point $\mathbf{x}^s = \mathbf{x}$, and find $\nabla \mathbf{h}_{\mathbf{x}^s}$.
3- Compute measurements information $\{\mathbf{y}_{\mathbf{z}}, \mathbf{Y}_{\mathbf{z}}\}$.
4- Add information and computed updated state $\{\hat{\mathbf{y}}, \hat{\mathbf{Y}}\}$.
5- Recover the moments $\{\hat{\mathbf{x}}, \hat{\mathbf{P}}\}$.

Algorithm 1: Nonlinear least-squares using information.

B. Robust Non-Linear Least-Squares

The main problem with least squares is its high sensitivity to outliers. This happens because the Gaussian has extremely small tails compared to most real measurement error distributions. A single outlier such as a correspondence error may affect one or a few of the observations, but it will usually leave all of the others unchanged. This locality is the whole basis of robustification. If we can decide which observations were affected, we can remove them and use the remaining observations for the parameter estimates as usual. Unfortunately, the definition of an outlier depends on the linearisation of $\mathbf{h}(\mathbf{x})$, and hence the choice of the linearisation point $\mathbf{x}^s$. Standard recursive estimators, such as the Extended Kalman Filter (EKF) and the Extended Information Filter (EIF), choose to linearise around the predicted state. Thus, in standard recursive estimators $\mathbf{x}^s = \hat{\mathbf{x}}$. This is a significant limitation, and we will instead allow the linearisation point to move during the estimation process as more information becomes available. This is beneficial because an accurate linearisation point will produce accurate linearised models, and hence accurate projection of uncertainty. Thus, it will improve our decisions on outliers and hence improve the robustification of the data fusion equations.

Several data association techniques can be applied to detect outliers in the data [4]. They include Nearest-Neighbor (NN), Maximum Likelihood (ML) formulation (or Individual Compatibility (IC) criterion) and the Joint Compatibility Branch and Bound algorithm (JCBB). In this section, we make data fusion decisions based on the IC criterion. Where measurements are tested individually and the decision of whether to include or exclude a measurement solely depends on the compatibility of individual measurements. First, at the *front-end* before data fusion takes place, and then at the *back-end* after data fusion has taken place. Notice that during estimation we should keep the state estimate $\mathbf{x}$ fixed, but we allow the linearisation point $\mathbf{x}^s$ to move. This is because a previously rejected observation can become an inlier given a new linearisation point, while a previously accepted observation can become an outlier.

Front-end gating utilises innovations vector ν which is defined as the difference between predicted measurements $\hat{\mathbf{z}}$ at a linearisation point $\mathbf{x}^s$ and actual measurements which are

currently have not been included yet $\mathbf{z}^{\text{off}}$, evaluated before data fusion takes place,

$$\nu = \mathbf{z}^{\text{off}} - \mathbf{h}(\mathbf{x}^s) - \nabla \mathbf{h}_{\mathbf{x}^s}(\mathbf{x} - \mathbf{x}^s).$$

Under the Gaussian assumption, an innovations gate g_ν can be defined as the Normalised Innovations Squared (NIS),

$$g_\nu = \nu^\top \mathbf{S}_\nu^{-1} \nu. \tag{10}$$

where, $\mathbf{S}_\nu$ is the innovations covariance which is given by,

$$\mathbf{S}_\nu = \nabla \mathbf{h}_{\mathbf{x}^s} \mathbf{P} \nabla \mathbf{h}_{\mathbf{x}^s}^\top + \mathbf{R}.$$

Under the hypothesis H_0 that the least-squares estimator is consistent, then the normalised innovations squared are Chi-squared $\mathcal{X}^2_m(\alpha)$ distributed in $m = \dim(g_\nu)$ degrees of freedom within a specified probability $1 - \alpha$. Thus, a confidence interval $[a_1, a_2]$ between which the innovations should lie can be constructed to test if the hypothesis H_0 that g_ν is indeed distributed as $\mathcal{X}^2_m(\alpha)$ should be accepted;

$$P(g_\nu \in [a_1, a_2] | H_0) = 1 - \alpha.$$

However, to detect outliers, checking that the gating function g_ν is lower than the upper bound a_2 is enough. Information by measurements passing the innovations gate is then added into the estimated model using simple addition operation,

$$\mathbf{y}^+ = \mathbf{y} + \mathbf{y}_{\mathbf{z}}^{\mathbf{off}(g_\nu < a_2)},$$
$$\mathbf{Y}^+ = \mathbf{Y} + \mathbf{Y}_{\mathbf{z}}^{\mathbf{off}(g_\nu < a_2)},$$
$$\mathbf{on} = \mathbf{off}(g_\nu < a_2) \cup \mathbf{on},$$
$$\mathbf{off} = \mathbf{off}(g_\nu > a_2).$$

Back-end gating utilises residuals vector $\mathbf{r}$ which is defined as the difference between predicted measurements $\hat{\mathbf{z}}$ at a linearisation point $\mathbf{x}^s$ and actual measurements that have been previously included $\mathbf{z}^{\text{on}}$, evaluated after data fusion takes place,

$$\mathbf{r} = \mathbf{z}^{\text{on}} - \mathbf{h}(\mathbf{x}^s) - \nabla \mathbf{h}_{\mathbf{x}^s}(\mathbf{x}^+ - \mathbf{x}^s).$$

Under the Gaussian assumption, a residual gate $g_{\mathbf{r}}$ can be defined as the Normalised Residuals Squared (NRS),

$$g_{\mathbf{r}} = \mathbf{r}^\top \mathbf{S}_{\mathbf{r}}^{-1} \mathbf{r}. \tag{11}$$

Where $\mathbf{S}_r$ is the residuals covariance which is given by,

$$\mathbf{S}_r = \nabla \mathbf{h}_{\mathbf{x}^s} \mathbf{P}^+ \nabla \mathbf{h}_{\mathbf{x}^s}^\top + \mathbf{R}.$$

A previously added measurement is considered an outlier given the most recent estimate $\hat{\mathbf{x}}^+$ if its residuals $g_{\mathbf{r}}$ is higher than a_2. Information by measurements failing the residuals is then subtracted from the estimated model using simple subtraction operation,

$$\hat{\mathbf{y}}^+ = \mathbf{y}^+ - \mathbf{y}_{\mathbf{z}}^{\text{on}(g_{\mathbf{r}} > a_2)},$$
$$\hat{\mathbf{Y}}^+ = \mathbf{Y}^+ - \mathbf{Y}_{\mathbf{z}}^{\text{on}(g_{\mathbf{r}} > a_2)},$$
$$\mathbf{off} = \mathbf{on}(g_{\mathbf{r}} > a_2) \cup \mathbf{off},$$
$$\mathbf{on} = \mathbf{on}(g_{\mathbf{r}} < a_2).$$

Finally, the posterior mean and covariance can be recovered by solving the following sparse system of equations,

$$\hat{\mathbf{Y}}^+ \hat{\mathbf{x}}^+ = \hat{\mathbf{y}}^+, \quad \hat{\mathbf{P}}^+ = \hat{\mathbf{Y}}^{+^{-1}},$$

which are known as the normal equations for the nonlinear least squares problem. Since the information matrix is typically sparse and positive definite[6], the nonlinear least-squares system can be solve efficiently using the Cholesky factorisation of the information matrix. Algorithm 2 summarises steps for robust data fusion using iterative least-squares.

When performing robust data fusion in the information form, the information matrix is inverted twice during each iteration. Since the information matrix is typically sparse, performing data fusion in the information form is cheaper than using the covariance form. Which also requires inverting the covariance matrix twice during data fusion, but its typically dense. Also, notice that the modifications to the information vector and matrix only concern those portions of the state appearing explicitly in the predicted measurements model. For problems where the state vector becomes very large but the individual measurements are functions of just a few states, the resultant information matrix is sparse and requires minimum operations to update given the measurements.

1- Initialise information vector and matrix $\{\mathbf{y}, \mathbf{Y}\}$.
2- Set linearisation point $\mathbf{x}^s = \mathbf{x}$, and find $\nabla \mathbf{h}_{\mathbf{x}^s}$.
3- Reset switch vectors $\{\mathbf{on}, \mathbf{off}\}$.
4- Compute measurements information $\{\mathbf{y}_{\mathbf{z}}, \mathbf{Y}_{\mathbf{z}}\}$.
5- For measurements $\mathbf{z}^{\mathbf{off}}$, compute gate g_ν.
6- Add information with $g_\nu < a_2$, update $\{\mathbf{on}, \mathbf{off}\}$.
7- For measurements $\mathbf{z}^{\mathbf{on}}$, compute gate $g_{\mathbf{r}}$.
8- Subtract information with $g_{\mathbf{r}} > a_2$, update $\{\mathbf{on}, \mathbf{off}\}$.
9- Recover the moments $\{\hat{\mathbf{x}}^+, \hat{\mathbf{P}}^+\}$.
10- Repeat from step 2 for T trials.

Algorithm 2: Robust nonlinear least-squares with innovations and residuals gating.

V. Smoothing and Mapping Using Robust Nonlinear Least-Squares

The following sections describe how algorithm 2 can be applied to solve the two estimation problems in sections II-A and II-B in relation to view-graph SLAM.

A. Robust Estimation of View-Graph Constraints

Given measurements $\mathbf{z}$ in (2) and measurements model $\mathbf{h}(\mathbf{x})$ in (3), an initial linearisation point $\mathbf{x}^s$ is required to solve for the relative states $\mathbf{x}$ in (1) and to evaluate the validation gates in (10) and (11). However, a good starting set of inlier

[6]Even though, the information matrix is positive definite by definition,

$$\mathbf{x}^\top \mathbf{Y} \mathbf{x} = \mathbf{E}\left[\left(\sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i} \log\, p(\mathbf{x})\right)^2\right] \succeq \mathbf{0},$$

additional tests are required to guarantee its numerical stability during information addition and subtraction steps. This guarantees that a Cholesky decomposition can be computed, and hence the information matrix can be inverted efficiently using Takahashi's inverse [17].

correspondences $\{\mathbf{m}^1_{1:N}, \ldots, \mathbf{m}^M_{1:N}\}$ are required for acceptable initialisation. One way to initialise relative poses is by using RANSAC with the essential matrix. Additionally, points selection and alignment using techniques, including features bucketing and non-maximal suppression, can be utilised to guarantee the accuracy of features localisation and epipolar geometry computations. Given that rotations can be computed from the essential matrix more accurately than translations, which can be only computed as unit vectors. We initialise relative rotations $\mathbf{a}_{1j}$ from the essential matrix. For translations $\mathbf{t}_{1j}$, we essentially need at least one scaled translation in order to maintain the map scale. This translation (and hence the scale) is then refined using the remaining frames in the bundle. Thus, we initialise the first relative translation $\mathbf{t}_{12}$, using roughly known initial relative geometry between the first two camera frames in the bundle[7]. Without this initial scale condition, translations can be only estimated up to a scale factor. Finally, inverse depth states are initialise using standard two-view triangulation.

The next step is to initialise the information matrix $\mathbf{Y}$ and then compute the information vector $\mathbf{y}$. One strategy to initialise the inverse depth terms of the information matrix is using the linearised projection of uncertainty from pixel-space to state-space. Yet another simplified rather fragile strategy, that does not require explicit projection of uncertainties, is to initialise a sparse information matrix and an information vector using,

$$\mathbf{Y} = \text{diag}([\mathbf{10^6}_{1\times 3}, \mathbf{10^6}_{1\times 3}, \mathbf{10^6}_{1\times 3}, \mathbf{0}, \ldots, \mathbf{0}]), \quad \mathbf{y} = \mathbf{Y}\mathbf{x}.$$

Thus, the information matrix dimensions are $(6M + N)\times(6M + N)$. Notice that the information terms concerning the reference camera pose $\{\sigma^{-2}_{\mathbf{t}_{11}}, \sigma^{-2}_{\mathbf{a}_{11}}\}$ are set to a large value $\mathbf{10^6}$. This implies very small initial uncertainty, and hence the reference frame is not allowed to move during estimation. Similarly, the information terms concerning the first scaled relative translation $\sigma^{-2}_{\mathbf{t}_{12}}$ are set to a large value. While the remaining relative motion and inverse depth terms are initialised as all zeros. This implies that no prior information is available at the time of initialisation, and the only source of pose and depth information is due to the measurements.

Finally, algorithm 2 is used to fuse inlier measurements and recover the state $\{\mathbf{x}, \mathbf{P}\}$ of the system. In this case, optimising for the normalised innovations squared and the normalised residuals squared is equivalent to optimising for the *normalised image projection error squared*. Notice that, during estimation, we use all the correspondences between the frames, and we do not utilise RANSAC iterations to prefilter outliers. This is based on the assumption that the initial poses by RANSAC and the essential matrix are inaccurate, and hence pose fixation could be encapsulated in the corresponding outliers information. Figure 3 illustrates the effect of information addition and subtraction on the map scale.

B. Robust View-Graph Estimation

Given measurements $\mathbf{z}$ in (6) and measurements model $\mathbf{h}(\mathbf{x})$ in (5), an initial linearisation point is needed to solve for the graph states $\mathbf{x}$ in (4) and to evaluate the validation gates in (10) and (11). In *unweighted graphs*, one way to initialise a pose estimate $\mathbf{x}$ robustly while accounting for outliers is from a spanning tree using RANSAC. However, given a *weighted graph*, a maximum spanning tree (MST) provides a viable initial solution. In this case, edge weights $\mathbf{w}_{ij}$ may represent the accuracy of the relative pose constraints in the graph. Various accuracy measures can be used, including the number of inlier matches between pair-wise geometries [11] and the trace of the estimated relative covariance matrix [12]. In the second case, the trace of updated relative pose covariance matrix is used, which can be computed from the information matrix estimated in the previous section by the relative pose optimiser. Hence, given the MST in a view-graph, the global poses can be initialised by accumulating measurements by its edges.

$$\mathbf{t}_1 = \mathbf{0}_{3\times 1}, \quad \mathbf{C}_1 = \mathbf{I}_{3\times 3},$$

$$\mathbf{C}_j = \mathbf{C}_i \mathbf{C}_i^{j\top}, \quad \{i, j\} \in \text{MST},$$

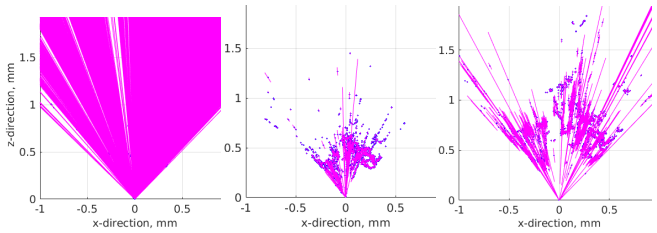
$$\mathbf{t}_j = \mathbf{t}_i + \mathbf{C}_i^j \mathbf{t}_{ij}, \quad \{i, j\} \in \text{MST}.$$

Again, we chose to keep all the constraints in the graph, and to initialise the global pose estimate from the MST in a weighted view-graph. This allows for a more efficient initialisation, since RANSAC iterations are not used. While outliers are detected and removed during estimation.

Given a graph initial state $\mathbf{x}$, the information vector and matrix are initialised as all zeros. This implies that no prior information is available at the time of initialisation, and the only source of pose information is due to the measurements. While the first frame is defined as the origin, and thus is initialised with large information.

$$\mathbf{Y} = \text{diag}([\mathbf{10^6}, \mathbf{0}, \ldots, \mathbf{0}]), \quad \mathbf{y} = \mathbf{Y}\mathbf{x}.$$

The information matrix has dimensions of $(6K \times 6K)$. The very large information at the origin implies that the first camera frame is not allowed to move away from the origin during estimation. Finally, algorithm 2 is used to fuse inlier measurements and recover the state $\{\mathbf{x}, \mathbf{P}\}$ of the system.



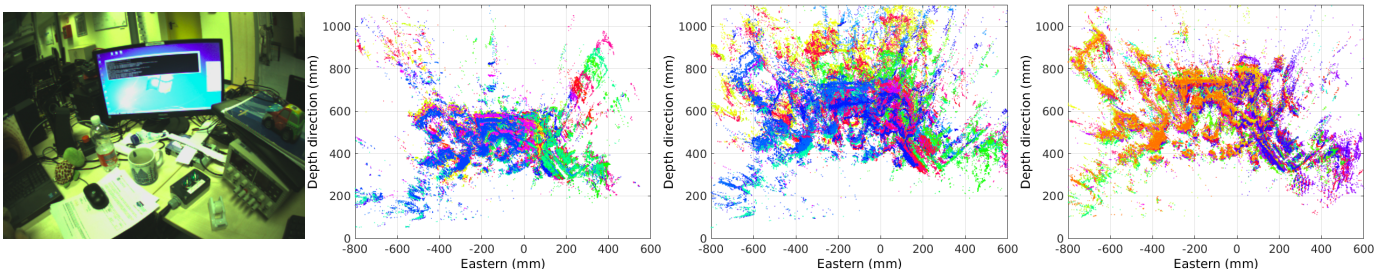
(a) $\{0, 114982\}$ (b) $\{114755, 227\}$ (c) $\{86001, 28981\}$

Fig. 3. A visual scan with inverse depth uncertainty for a number of iterations as follows: (a) At initialisation. (b) After first addition step with $\alpha = 95\%$. (c) After last subtraction step with $\alpha = 95\%$. Also shown the number of constraints as $\{\mathbf{on}, \mathbf{off}\}$. The figure shows the effect of constraints addition and subtraction on depth uncertainty.

VI. Experiments

The following experimental results used a sequence of images obtained from iCub and then processed using an Intel Core i5 CPU@1.90GHz×4.

[7]For the case of our robot, iCub, $\mathbf{t}_{12}$ equals roughly $[68, 0, 0]^\top$ millimeter. This provides an additional constraint on the scale of the map.



(a) Sample image. (b) Pair-wise kinematics. (c) RANSAC, Essential matrix and LM. (d) Algorithm 2.

Fig. 4. Two-view visual scans (top view) obtained using three different methods. Since a global pose estimate is not yet available, scans were projected into a common global frame using kinematics.

A. Vision System

Although accurate localisation does not require extraction of dense map landmarks, good data association and quality image features are required for better performance. In this implementation, FAST corners [18] were extracted in one of the K image frames in the graph and then tracked in the consecutive frames using a pyramidal implementation of Lucas-Kanade optical flow tracker [19]. Tracking was carried out based on a pre-defined graph connectivity (redundancy is defined by the number of view constraints M, in section II-A). This reduces the cost of tracking, since corners extracted in one frame were only tracked in the following $M-1$ frames. A link with less than a pre-defined number of point tracks (50 in this paper) was removed from the graph.

B. Relative Two-View Results

Figure 4 shows visual scans computed using two-view constraints (M=2). The figure shows multiple visual scans projected into the global common coordinates using kinematics odometry. The figure demonstrates the robustness of our method in comparison to using relative stereo kinematics alone or using RANSAC (500+ iterations) and Levenberg-Marquardt (LM) nonlinear least-squares method. Where, the variation in the map scale was more significant when RANSAC and LM were used. This was because of the ambiguity of relative motion computations when only two view constraints were given. This ambiguity, however, was better resolved when the relative kinematics were used. Even though every scan was computed from two-view constraints, which is often considered as a fragile implementation in computer vision, the robustness of our least-squares method is demonstrated by the consistency of scan scales in the map.

C. Global Pose and Map Results

In this section, a graph with 250 views ($K = 250$) and 19 paths ($M = 20$) is assumed. In this case, the most possible number of constraints equals $\frac{K(K-1)}{2}$. Figure 5 shows graph optimisation results using Algorithm 2. Using total of 4311 constraints, graph optimisation took on average 19 seconds. The figure shows the classification of graph constraints into inliers and outliers, and the estimated camera poses (where a measurements-noise matrix $\mathbf{R}$ with standard deviations of 0.2 degrees and 1 mm was assumed when evaluating the gates in (10) and (11)). Also, the figure compares the estimated cameras orientation against MST and the method in [14] and shows relative stereo anglular variations while moving the robot head. These relative angular variations from those computed from kinematics during head movements illustrate the existence of mechanical errors and hardware delays (which may case frame drops) between the robot eyes.

Figure 6 shows the reconstruction of the map after projecting the estimated scans by using their corresponding global poses. The figure compares the map reconstructed using graph pose estimates with that reconstructed using poses from kinematics. The figure illustrates the benefit of using vision information along with robust estimation in the alignment of visual scans, where the largest alignment error component is along the optical rays emanating from the reference cameras in the depth direction. This error can be reduced by including more M-view constraints (larger M) and applying more measurements switching iterations.

VII. Conclusions

This paper has presented a new robust nonlinear least-squares method to view-graph SLAM. The method was applied to SLAM using iCub, which includes a *non-rigid* stereo system with a short baseline. Robustness in this case is essential to reduce bias introduced by outliers in the data and the observability of the system due to rotations being the dominant motion component. The least squares method presented was proved robust given rough initialisation of the cameras and the map states.

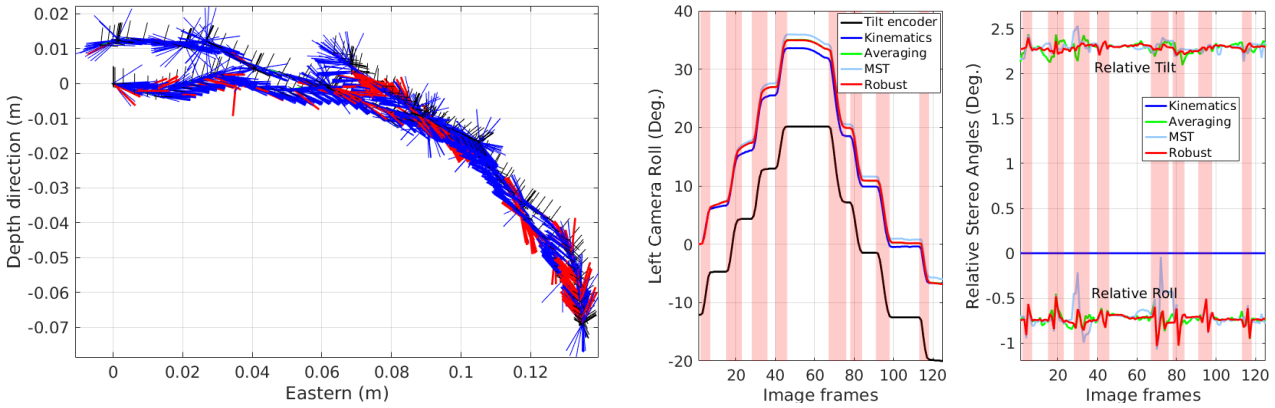
Future work will examine the scalability and efficiency of the method with the increased dimensions of the problem. We will also investigate real-time capabilities of the algorithms presented and look at methods for depth maps fusion and update to integrate the generated scans into a globally concise map. Yet another potential direction of improvement is to look at replacing the re-projection error cost function in (10) and (11) with a photometric one. This allows for denser and possibly more accurate mapping results.

Acknowledgment

This project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement num. FP7-ICT-611909 (KoroiBot).

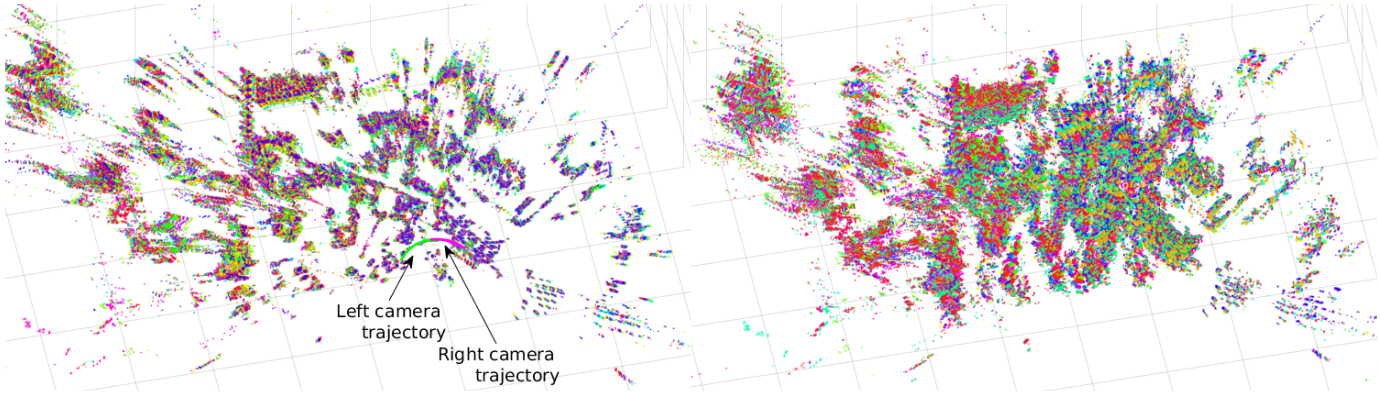
The authors would like to thank Tim Bailey[8] for his discussions and ideas on robustification of nonlinear least-squares.

[8]http://www-personal.acfr.usyd.edu.au/tbailey/



(a) View-graph. (b) Monocular and stereo angles.

Fig. 5. Results of view-graph optimisation with totals $\{\mathbf{on}, \mathbf{off}\} = \{3042, 1269\}$, where (a) shows view-graph constraints with accepted in blue and rejected in red. Also, left and right camera coordinate frames are shown in black. (b) shows angular differences between estimated orientation of the left camera using rotation averaging in [14], MST, and algorithm 2 and computed orientation from kinematics. Also shown the relative tilt and roll angles between stereo pairs in the sequence. Notice that the constraints used for both our method and rotation averaging were processed according to section V-A.



(a) Global registration using view-graph estimate. (b) Global registration using kinematics.

Fig. 6. Results of global registration of estimated visual scans, using M=20 and K=250. (a) Shows scans registered using view-graph estimates, and (b) Shows scans registered using iCub kinematis odometry.

REFERENCES

[1] J. Engel, J. Stuckler and D. Cremers, "Large-scale direct SLAM with stereo cameras," Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, Hamburg, pp. 1935-1942.

[2] Sameer Agarwal and Keir Mierle and Others, "Ceres Solver", http://ceres-solver.org.

[3] Kummerle, R.; Grisetti, G.; Strasdat, H.; Konolige, K.; Burgard, W., "g2o: A general framework for graph optimization," in Robotics and Automation (ICRA), 2011 IEEE International Conference on , vol., no., pp.3607-3613, 9-13 May 2011.

[4] Kaess, M.; Ranganathan, A.; Dellaert, F., "iSAM: Incremental Smoothing and Mapping," in Robotics, IEEE Transactions on , vol.24, no.6, pp.1365-1378, Dec. 2008.

[5] Nister, D., "An efficient solution to the five-point relative pose problem," in Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.26, no.6, pp.756-770, June 2004.

[6] Richard Hartley and Andrew Zisserman. 2003. Multiple View Geometry in Computer Vision (2nd edition). Cambridge University Press, New York, NY, USA.

[7] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon, "Bundle Adjustment - A Modern Synthesis". In Proceedings of the International Workshop on Vision Algorithms: Theory and Practice (ICCV '99).

[8] Govindu, Venu Madhav, "Robustness in motion averaging". Computer Vision – ACCV 2006: 7th Asian Conference on Computer Vision, Hyderabad, India. Proceedings, Part II, page 457–466.

[9] Jian, Yong-Dian, Doru C. Balcan, and Frank Dellaert. "Generalized subgraph preconditioners for large-scale bundle adjustment." Outdoor and Large-Scale Real-World Scene Analysis. Springer Berlin Heidelberg, 2012. 131-150.

[10] Olsson, C.; Eriksson, A.; Hartley, R., "Outlier removal using duality," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on , vol., no., pp.1450-1457, 13-18 June 2010.

[11] Carl Olsson; Olof Enqvist. 2011. Stable structure from motion for unordered image collections. In Proceedings of the 17th Scandinavian conference on Image analysis (SCIA'11), Anders Heyden and Fredrik Kahl (Eds.). Springer-Verlag, Berlin, Heidelberg, 524-535.

[12] Snavely, N.; Seitz, S.M.; Szeliski, R., "Skeletal graphs for efficient structure from motion," in Computer Vision and Pattern Recognition, CVPR 2008. IEEE Conference on , vol., no., pp.1-8, 23-28 June 2008.

[13] Crandall, D.; Owens, A.; Snavely, N.; Huttenlocher, D., "Discrete-continuous optimization for large-scale structure from motion," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on , vol., no., pp.3001-3008, 20-25 June 2011.

[14] Chatterjee, A.; Govindu, V.M., "Efficient and Robust Large-Scale Rotation Averaging," in Computer Vision (ICCV), 2013 IEEE International Conference on , vol., no., pp.521-528, 1-8 Dec. 2013.

[15] Kyle Wilson; Noah Snavely, "Robust Global Translations with 1DSfM", Computer Vision - ECCV, 13th European Conference, Proceedings, Part III, pp.61-75, Zurich, Switzerland, 2014.

[16] Bondy, John-Adrian and Murty, U. S. R., "Graph theory", Graduate texts in mathematics, Springer, New York, London, 2007.

[17] J Vanhatalo, A Vehtari, "Modelling local and global phenomena with sparse Gaussian processes", Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, 2008.

[18] Edward Rosten and Tom Drummond. 2006. "Machine learning for high-speed corner detection". In Proceedings of the 9th European conference on Computer Vision - Volume Part I (ECCV'06), Springer-Verlag, Berlin, Heidelberg, 430-443.

[19] Jean-Yves Bouguet. "Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm". Technical report, Intel Corporation Microprocessor Research Labs, 2000.