# ROBUST LOOP CLOSURES FOR SCENE RECONSTRUCTION BY COMBINING ODOMETRY AND VISUAL CORRESPONDENCES

*Zakaria Laskar, Sami Huttunen, Daniel Herrera C., Esa Rahtu and Juho Kannala*[†]

University of Oulu, Finland [†]Aalto University, Finland

## ABSTRACT

Given an image sequence and odometry from a moving camera, we propose a batch-based approach for robust reconstruction of scene structure and camera motion. A key part of our method is robust loop closure disambiguation. First, a structure-from-motion pipeline is used to get a set of candidate feature correspondences and the respective triangulated 3D landmarks. Thereafter, the compatibility of each correspondence constraint and the odometry is evaluated in a bundle-adjustment optimization, where only compatible constraints affect. Our approach is evaluated using data from a Google Tango device. The results show that it produces better reconstructions than the device's built-in software or a state-of-the-art pose-graph formulation.

***Index Terms***— bundle-adjustment, loop closures

## 1. INTRODUCTION

Three-dimensional scene reconstruction from multiple images is a popular research topic in computer vision, and there has been significant progress both in large-scale structure from motion (SfM) [1] and in real-time simultaneous localization and mapping (SLAM) [2] during the recent years.

However, despite all the progress, there are still major challenges in purely image-based reconstruction techniques, especially in indoor environments. Such challenges include lack of texture and structure in the scene (homogeneous textureless surfaces are common indoors) and degenerate motion sequences (e.g. triangulation requires translation and pure rotational motion often breaks camera-based tracking [3]). The problems due to lack of texture are further emphasized when narrow field-of-view cameras are used indoors where the scene surfaces are close to the camera and therefore the so-called *aperture problem* may occur more often than outdoors.

However, luckily, by combining visual loop closure detections and visual odometry [4], or other kind of odometry (e.g. [5, 6]), it may be possible to resolve ambiguous cases and discard the false positive loop closures, which are not compatible with the odometry or correct loop closures [7, 8]. Similar ideas are also used in the recent work [9] which focuses on dense high-quality surface reconstructions in indoor environments. Their approach gives precise models but it is not as scalable as large-scale structure-from-motion pipelines used for large outdoor datasets [1]. In fact, robust and scalable indoor reconstruction approaches, which could be ubiquitously used by amateurs with off-the-shelf hardware, have not yet emerged.

In this paper, we propose a method for robust scene reconstruction based on an image sequence, captured by a moving monocular camera, and the corresponding odometry information. That is, the input to our method are the image frames, their time stamps and a rough continuous odometry estimate of the camera trajectory. In addition, we assume that we have some information of the absolute length scale of the camera trajectory. For example, if the odometry track does not suffer from scale drift, it allows to define the common scale of the reconstruction, even if the available visual correspondences would be limited to several non-overlapping clusters of views. In our experiments we used Google Tango device which provides both depth maps and color images (so called RGB-D data) and also relatively good and robust odometry, which is computed by combining visual and inertial sensor information[1]. Unlike some previous methods [7, 8, 10], we do not use a strict pose graph formulation but explicitly model all visual correspondences in a bundle-adjustment like formulation [11, 12]. Thus, we can utilize and verify all available visual correspondences and correct the association errors of the structure-from-motion pipeline, whereas in many pose graph formulations the candidate loop closure constraints are utilized or discarded at image level.

## 2. PROPOSED SYSTEM

This section describes the different components of our proposed system. An overview is shown in Fig. 1. In our case, the input to the system is an RGB-D image sequence and odometry and the output is globally consistent 3D model with camera poses. However, depth maps are not absolutely necessary if the odometry has a consistent drift-free scale (cf. Sec. 2.4).

[1] https://developers.google.com/project-tango/

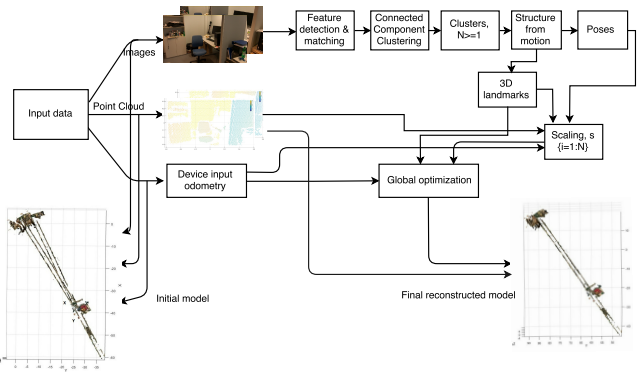



**Fig. 1**: An overview of our proposed system. Images, corresponding point cloud, and device odometry are input to our system. Our system corrects drift using triangulated 3D landmarks from SfM, and mitigates the effect of repetitive structures in the global optimization stage to give the final reconstructed model.

### 2.1. Feature detection and matching

Visual features are used to match images. Following [15, 13], SIFT [14] features and descriptors are used along with geometric filtering to generate visual correspondences.

### 2.2. Connected component clustering

Using the visual correspondences from the previous stage, an adjacency matrix is formed. One or more non-overlapping view-clusters are obtained by clustering the adjacency matrix using connected component clustering. Each view-cluster contains images that have been matched to at-least one other image in the cluster.

### 2.3. Structure from motion for view-clusters

Each view cluster is fed to a SfM pipeline [15]. The SfM pipeline triangulates each of the mutually observed features by the cameras in the cluster to a unique point in 3D world coordinates. If $\mathbf{x}_f$ is a triangulated 3D point, it is related to its observed image coordinate (sub-pixel position of feature $f$) $\mathbf{n}_c^f = [n_x\ n_y]^{\mathrm{T}}$ in camera c with pose estimates $[\mathbf{R}_c|\mathbf{t}_c]$ (from SfM) through the following equation :

$$\mathbf{p}_c^f = \mathbf{K}(\mathbf{R}_c\mathbf{x}_f + \mathbf{t}_c) \tag{1}$$

Here, $\mathbf{p}_c^f = d[n_x\ n_y\ 1]^{\mathrm{T}}$ will be defined as the feature measurement, and $d$ is the depth of the feature $f$ in the coordinate frame of camera $c$. $\mathbf{K}$ is the known camera calibration matrix. We define a function $\mathcal{P} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ such that $\mathcal{P}(\mathbf{p}) = \mathbf{n}$, $d(n_x, n_y, 1)^{\mathrm{T}} \rightarrow (n_x, n_y)^{\mathrm{T}}$. Only feature measurements and the observed image coordinates of the triangulated features are used in the final global optimization step.

Although we have RGB-D images as input, we use only RGB image features triangulated via SfM as the input depth maps are sparse and majority of pixels do not have a depth.

### 2.4. Scaling

The view clusters reconstructed via SfM may have different scales and in order to get a consistent initialization for our bundle-adjustment, we need to set the scales consistently. To set the scales we may utilize either the scale of depth maps or the scale of odometry between nearby cameras. In our experiments we utilized depth maps as follows.

Let $\mathbf{v}$ and $\mathbf{u}$ represent vectors containing the depth values obtained from SfM pipeline and input RGB-D data, respectively. Note that $\mathbf{v}$ holds the depth estimates of only those pixels for which there exists a depth value in the input frame data. The scaling factor is then computed as :

$$s = (\mathbf{u}^{\mathrm{T}}\mathbf{u})^{-1}\mathbf{u}^{\mathrm{T}}\mathbf{v} \tag{2}$$

Scaling removes the projective ambiguity associated with the structure from motion estimation. The scaled depth acts as a good initial estimate for the global bundle-adjustment process in the next stage.

### 2.5. Global optimization via bundle-adjustment

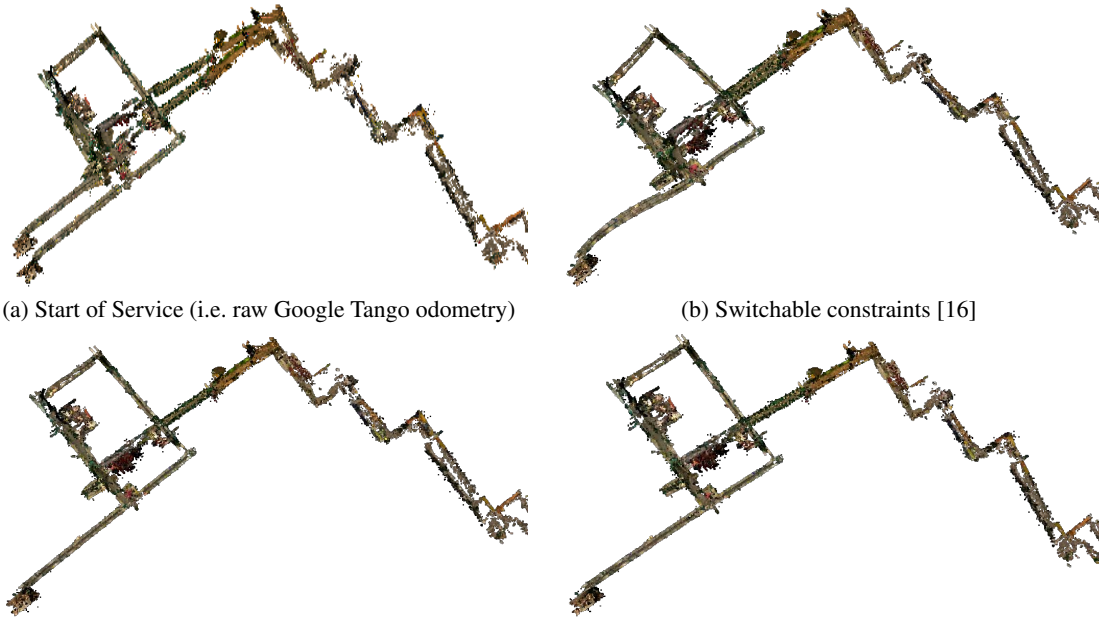
Let $C = \{c_1, c_2 ... c_T\}$ be the input camera sequence and $F$ denote the set of visual features obtained from Section 2.1. Each feature $f \in F$ is observed by a set of cameras $C_f = \{c_i | i \in \{1..T\}\}$ . Clustering the cameras $c \in C_f$ based on their time-stamp or image index gives us $M(\geq 1)$ non-overlapping camera-clusters $C_f^j \subseteq C_f$, $j$ =1...$M$, where $j$ represents a continuous stretch of time when a 3D landmark $\mathbf{x}_f$ was observed by a certain subset of the cameras $C_f$ causing the respective feature projections. Due to repetitive structure/pattern in space or spurious match, camera-clusters may be viewing a spatially different but visually similar 3D point altogether. As such we associate a unique 3D point $\mathbf{x}_f^j$ with each camera-cluster $C_f^j$. By putting the feature measurement $\mathbf{p}_c^f$ and the pose estimate $[\hat{\mathbf{R}}_c|\hat{\mathbf{t}}_c]$ from input odometry of any one of the cameras $c \in C_f^j$ in equation (1), we obtain an initial estimate of $\mathbf{x}_f^j$. We now introduce the set $H_f = \{\{\mathbf{x}_f^1, C_f^1\}, \{\mathbf{x}_f^2, C_f^2\} ... \{\mathbf{x}_f^M, C_f^M\}\}$ to represent the various 3D landmarks with similar visual description and its conceiving camera-cluster.

Each 3D point $\mathbf{x}_f^j$ is only constrained by its point-projections on the cameras $c \in C_f^j$. The constraint is modelled as the 2D reprojection error, which is the squared distance between a 3D point $\mathbf{x}_f^j$'s projection $\hat{\mathbf{n}}$ and the observed image coordinates $\mathbf{n}_c^f$ in image $c$ using the current pose estimates $[\mathbf{R}_c|\mathbf{t}_c]$ :

$$E_{2D,f,j,c} = ||\mathbf{n}_c^f - \mathcal{P}(\mathbf{K}(\mathbf{R}_c\mathbf{x}_f^j + \mathbf{t}_c))||^2 \tag{3}$$

However, some of the 3D points in $H_f$ may represent the same landmark in space. In order to merge the potential loop closure landmarks we use a 3D alignment error term. The 3D alignment error is the squared distance between two 3D world coordinates m = $\{\mathbf{x}_f^k, \mathbf{x}_f^l\}$:

$$E_{3D,f,m} = ||\mathbf{x}_f^k - \mathbf{x}_f^l||^2 \tag{4}$$



(a) Start of Service (i.e. raw Google Tango odometry)

(b) Switchable constraints [16]

(c) Area Learning (Google Tango)

(d) Our approach

**Fig. 2**: Campus dataset. The points of the Google Tango depth maps are illustrated from above. The camera trajectory is almost 600 meters. All the three methods (b)-(d) were able to limit the drift of the odometry (a) by imposing loop closure constraints. However, in terms of global consistency, our approach outperforms the other two as can be seen in a detailed view of a particular region of the map in Figure 3.

.

Additionally, we add a smoothness regularizer in the form of the relative odometry error term. It is the error in relative pose between sequential images using current pose estimates $[\mathbf{R}|\mathbf{t}]$ and the pose estimates from input odometry $[\hat{\mathbf{R}}|\hat{\mathbf{t}}]$:

$$\begin{aligned} E_{\Delta R,i} &= ||\tau(\tau(\mathbf{R}_{i-1}, \mathbf{R}_i), \tau(\hat{\mathbf{R}}_{i-1}, \hat{\mathbf{R}}_i))||^2 \\ E_{\Delta t,i} &= ||(\gamma(\mathbf{t}_{i-1}, \mathbf{t}_i) - \gamma(\hat{\mathbf{t}}_{i-1}, \hat{\mathbf{t}}_i))||^2 \end{aligned} \tag{5}$$

where $\tau(\mathbf{R}_l, \mathbf{R}_j) = \mathbf{R}_j \mathbf{R}_l^{\mathrm{T}}$, and, $\gamma(\mathbf{t}_l, \mathbf{t}_j) = \mathbf{t}_j - \tau(\mathbf{R}_l, \mathbf{R}_j)\mathbf{t}_l$.

The final global optimization is formulated as follows :

$$\begin{aligned} \underset{\mathbf{x},\mathbf{r},\mathbf{t},w}{\operatorname{argmin}} & \sum_{f \in F} \sum_{j=1}^{M} \sum_{c \in C_f^j} \frac{E_{2D,f,j,c}}{\sigma_p^2} + \sum_{p=2}^{T} \frac{E_{\Delta R,p}}{\sigma_r^2} + \sum_{p=2}^{T} \frac{E_{\Delta t,p}}{\sigma_t^2} \\ & + \sum_{f \in F} \sum_{m \in \varphi_f} \frac{w_m . E_{3D,f,m}}{\sigma_a^2} + \sum_{f \in F} \sum_{m \in \varphi_f} \frac{||1 - w_m||^2}{\sigma_c^2} \end{aligned} \tag{6}$$

Here, $\mathbf{r}$ denotes the angle-axis representations of rotations $\mathbf{R}$. We minimize over $\mathbf{r}$, but while calculating errors in the observation model, we transform back from $\mathbf{r}$ to $\mathbf{R}$. As the errors have different units, they are weighted with the inverse of corresponding measurement variances $(\sigma_p^2, \sigma_r^2, \sigma_t^2, \sigma_a^2, \sigma_c^2)$. The set $\varphi_f = \{\{\mathbf{x}_f^1, \mathbf{x}_f^2\}, ., \{\mathbf{x}_f^2, \mathbf{x}_f^3\}, ..\{\mathbf{x}_f^{M-1}, \mathbf{x}_f^M\}\}$ contains all pairwise combinations of the 3D points in the set $H_f$. $w_m \in [0, 1]$ are switch variables [16, 7] associated with potential erroneous constraints. When associated with an erroneous constraint the switch variable drives its value to zero. As a result the erroneous constraints stops to have any impact on the optimization. Particularly, in our case, the switch variables restrict the minimization of 3D alignment error $E_{3D}$ between 3D point pairs in $\varphi_f$ when the pair represents different 3D points in space. The last term in equation (6) tries to constraint the switch variables at their initial values ($w_m$ = 1). The objective in equation (6) is minimized using standard non-linear least-squares optimization with trust-region algorithm [17].

Unlike pose-graph optimization methods which use relative pose between non-sequential cameras as loop closure constraint, we try to minimize the distance between 3D landmarks ($E_{3D}$). As the 3D landmarks are constrained by their observed image coordinates ($E_{2D}$), we are able to maintain structural consistency in regions of loop closure. Additionally, using relative odometry constraints ($E_{\Delta R}, E_{\Delta t}$) along with multi-view constraints ($E_{2D}$), we are able to mitigate the effect of erroneous visual correspondences and produce a globally consistent map.

## 3. RESULTS AND DISCUSSION

As there are no publicly available Google Tango datasets for evaluation, we captured three different indoor datasets by ourselves: Apartment, University and Metro Station (see supplemental). All the datasets contain frequently occurring repetitive patterns or structures. Table 1 gives a brief idea about the datasets.

We compare the performance of our approach with the Area Learning mode of Google Tango, which performs simul-

(a) Area Learning (b) Switchable Constraints (c) Our approach

**Fig. 3**: Detailed views of a region from reconstructions of Campus dataset.

(a) Start of service (b) Switchable Constraints (c) Area Learning (d) Our approach

**Fig. 4**: Reconstruction of Apartment dataset. (a) Raw odometry (Start of Service mode of Google Tango). (b) Reconstruction by Switchable Constraints. (c) Reconstruction by Area Learning functionality of Google Tango. (d) Reconstruction from our approach. Visual appearance of the walls and boundaries of different objects in the apartment in the three reconstructed models clearly demonstrate how our approach is able to correct drift in odometry while maintaining the consistency in structure.

taneous localization and mapping (cf. Fig. 2). We also compare with switchable constraints formulation, a state of the art pose-graph method[16], which does not explicitly reconstruct and adjust 3D feature points but is able to fuse conflicting pairwise relative pose constraints. As ground truth data is not available, the performance of our approach is evaluated through visual inspection of the reconstructed 3D models. A detailed view of some of the regions in a map are presented in respective figures. A more thorough discussion of the results is given in the following subsections:

**Apartment**: The Apartment dataset was collected in a 44 $m^2$ area apartment with a camera trajectory of 20 meters. Inside the apartment is a dense setting of objects and repetitive patterns across the walls. The results in Figure 4 demonstrates that our approach can obtain structural consistency with high fidelity.

**Campus**: The Campus dataset was collected in a university. The total trajectory is 170 meters in the second floor and 400 meters in the first floor. The results in Figure 2 show that all the three methods were able to limit drift by imposing loop closures, but only our approach could maintain the structural consistency as seen in Figure 3.

**Metro Station**: Due to lack of space our results with the Metro Station dataset can not be illustrated here (see supplemental) but these results are consistent with the others and confirm the observations made with the other datasets illustrated in Figures 2, 3 and 4.

**Table 1**: Dataset table.

| Dataset Name | frames | Trajectory length (m) |
|---|---|---|
| Apartment | 892 | 20 |
| Campus | 4021 | 588 |
| Metro station | 812 | 180 |

## 4. CONCLUSION

We presented an approach for indoor reconstruction from RGB-D images and odometry. The key idea is to include structural information with odometry in a single global bundle-adjustment process. The formulation makes the system highly robust to erroneous loop closures and geometric inconsistency as the structural information adds strict constraints to the global optimization.

Experimental results demonstrate that our approach performs significantly better in producing globally consistent 3D model of the map than the built-in SLAM software in Google Tango tablet and a state of the art pose-graph optimization method [16].

## 5. REFERENCES


[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M. Seitz, and Richard Szeliski, "Building Rome in a day," *Commun. ACM*, vol. 54, no. 10, pp. 105–112, Oct. 2011.

[2] Jakob Engel, Thomas Schöps, and Daniel Cremers, "LSD-SLAM: large-scale direct monocular SLAM," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., vol. 8690 of *Lecture Notes in Computer Science*, pp. 834–849. Springer International Publishing, 2014.

[3] Daniel Herrera C., Kihwan Kim, Juho Kannala, Kari Pulli, and Janne Heikkilä, "DT-SLAM: deferred triangulation for robust slam deferred triangulation for robust slam," in *3D Vision (3DV), 2014 2nd International Conference on*. IEEE, 2014, vol. 1, pp. 609–616.

[4] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, June 2004, vol. 1, pp. I–652–I–659 Vol.1.

[5] Mingyang Li, Byung Hyung Kim, and Anastasios I. Mourikis, "Real-time motion tracking on a cellphone using inertial sensing and a rolling-shutter camera," in *IEEE International Conference on Robotics and Automation*, 2013.

[6] Christian Kerl, Jurgen Sturm, and Daniel Cremers, "Robust odometry estimation for RGB-D cameras," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3748–3754.

[7] Niko Sünderhauf and Peter Protzel, "Towards a robust back-end for pose graph slam," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1254–1261.

[8] Niko Sünderhauf and Peter Protzel, "Switchable constraints vs. max-mixture models vs. rrr-a comparison of three approaches to robust pose graph slam," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 5198–5203.

[9] Sungjoon Choi, Qian-Yi Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, June 2015, pp. 5556–5565.

[10] Gim Hee Lee, Friedrich Fraundorfer, and Marc Pollefeys, "Robust pose-graph loop-closures with expectation-maximization," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 556–563.

[11] Christopher Zach, "Robust bundle adjustment revisited," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., vol. 8693 of *Lecture Notes in Computer Science*, pp. 772–787. Springer International Publishing, 2014.

[12] Juho Kannala, Sami S. Brandt, and Janne Heikkilä, "Measuring and modelling sewer pipes from video," *Mach. Vis. Appl.*, vol. 19, no. 2, pp. 73–83, 2008.

[13] K.K.S. Bhat, Juho Kannala, and Janne Heikkilä, "3D point representation for pose estimation: Accelerated SIFT vs ORB," in *Image Analysis*, Rasmus R. Paulsen and Kim S. Pedersen, Eds., vol. 9127 of *Lecture Notes in Computer Science*, pp. 79–91. Springer International Publishing, 2015.

[14] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110.

[15] Pierre Moulon, Pascal Monasse, Renaud Marlet, and Others, "Openmvg," .

[16] N. Sünderhauf and P. Protzel, "Switchable constraints for robust pose graph slam," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, Oct 2012, pp. 1879–1884.

[17] Yuying Li, "Centering, trust region, reflective techniques for nonlinear minimization subject to bounds," Tech. Rep., Ithaca, NY, USA, 1993.