Published in Proc. of IEEE Intl. Conf. on Intelligent Robots and Systems (IROS), 2011. DOI: http://dx.doi.org/10.1109/IROS.2011.6094921

BRIEF-Gist – Closing the Loop by Simple Means

Niko Sünderhauf and Peter Protzel

***Abstract*— The ability to recognize known places is an essential competence of any intelligent system that operates autonomously over longer periods of time. Approaches that rely on the visual appearance of distinct scenes have recently been developed and applied to large scale SLAM scenarios. FAB-Map is maybe the most successful of these systems.**

Our paper proposes BRIEF-Gist, a very simplistic appearance-based place recognition system based on the BRIEF descriptor. BRIEF-Gist is much more easy to implement and more efficient compared to recent approaches like FAB-Map. Despite its simplicity, we can show that it performs comparably well as a front-end for large scale SLAM. We benchmark our approach using two standard datasets and perform SLAM on the 66 km long urban St. Lucia dataset.

I. Introduction

Modern SLAM systems are typically based on the efficient optimization of probabilistic constraint or factor graphs. These systems are generally divided into a back-end and front-end [8]. The back-end contains the optimizer that builds and maintains a map by finding an optimal solution to the robot's trajectory and the landmark positions given the constraints constructed by the front-end. This front-end is responsible for data association in general and, in the context of pose-only SLAM, place recognition in particular.

Reliable place recognition is a hard problem, especially in large-scale environments. Repetitive structure and sensory ambiguity constitute severe challenges for any place recognition system. As optimization based back-ends for SLAM like iSAM [7], Sparse Pose Adjustment [8], iSAM2 [6], or g2o [10] are not robust against outliers, even a single wrong loop closure will result in a catastrophic failure of the mapping process.

Recent developments in appearance-based place recognition therefore aimed at reaching a high recall rate at 100% precision, i.e. they concentrated on preventing false positives. This of course leads to computationally involved, very complex systems.

In parallel work, we developed a robust formulation to pose graph SLAM that allows the optimizer in the back-end to identify and reject wrong loop closures. This can be understood as enabling the back-end to take back any data association decision of the front-end. Given this robust back-end, the need of reaching a precision of 100% during the data association (i.e. place recognition) process is eliminated. The place recognition system in the front-end can therefore be kept simple and focused on a high recall rate, as a reasonable number of false positive loop closures is acceptable.

Our paper proposes BRIEF-Gist, an appearance-based place recognition system that builds upon the BRIEF descriptor by Calonder et al. [3]. We evaluate BRIEF-Gist and conclude that our approach is suitable to perform place recognition in large scale scenarios, despite its simplicity regarding implementation and computational demands.

The authors are with the Department of Electrical Engineering and Information Technology, Chemnitz University of Technology, 09111 Chemnitz, Germany. {niko.suenderhauf, peter.protzel}@etit.tu-chemnitz.de

II. Related Work

A. Appearance-Based Place Recognition

Important work towards appearance-based place recognition has been conducted by Sivic and Zisserman [22] who borrowed ideas from text retrieval systems and introduced the concept of the so called *visual vocabulary*. The idea was later extended to *vocabulary trees* by Nister and Stewenius [16], allowing to efficiently use large vocabularies. Schindler et al. [21] demonstrated city-scale place recognition using these tree structures.

FAB-Map [4] is a probabilistic appearance-based approach to place recognition. It builds on a visual vocabulary learned from SURF descriptors [1]. A Chow Liu tree is used to approximate the probability distribution over these visual words and the correlations between them. This allows the system to robustly recognize known places despite visual ambiguity. FAB-Map 2.0 [5] has been applied to a 1000 km dataset and achieved a recall of 3.1% at 100% precision (14.3% at 90 % precision respectively).

Recently, Cadena et al. [2] combined appearance-based place recognition with Conditional Random Fields to filter out mismatches caused by visual ambiguity between spatially distinct places.

Maddern et al. [12] report an improvement to the robustness of FAB-Map by incorporating odometric information into the place recognition process.

The methods mentioned above describe the appearance of a scene through distinct landmarks (feature points) and their descriptors. Another strategy is to use so called *holistic* descriptors, i.e. descriptors that describe the appearance of the complete scene and not of single points in it. The idea of a holistic scene descriptor is not new and was e.g. examined by Oliva and Torralba [18] [17] with the introduction of the *Gist* descriptor. This global image descriptor is built from the responses of steerable filters at different orientations and scales. More recently, [14] demonstrated place recognition using the Gist descriptor on panoramic images in an urban environment.

B. The BRIEF Descriptor

BRIEF (*Binary Robust Independent Elementary Features*) has been introduced as an efficient descriptor for feature points (or keypoints) by Calonder et al. [3]. It was found to be superior to the established SIFT [11] or SURF [1] descriptors, both in recognition performance and runtime behaviour.

The BRIEF-descriptor is a bit-vector (e.g. of length 256) that is built by simple binary tests on a subset of the pixels surrounding the keypoint center. Calonder et al. [3] suggest using a simple comparison of pixel intensity values: For a descriptor of length n (e.g. $n = 256$), n pixel-pairs $(p_{k,1}, p_{k,2})$ are chosen in the local neighborhood (e.g. 48×48) of the keypoint center. The k-th bit in the descriptor is set to 1 if $p_{k,1} < p_{k,2}$ and set to 0 otherwise. This way, the descriptor can be built very efficiently. Notice that the same neighboring pixels will be chosen for all descriptors.

Comparing two descriptors D_1 and D_2, i.e. determining their similarity, can be performed very efficiently using the Hamming distance (which is the L_1 norm). As the descriptors are simple bit-vectors, their Hamming distance can be calculated by

$$\|D_1 - D_2\|_H = \mathrm{bitsum}(D_1 \oplus D_2) \tag{1}$$

where $\oplus$ is the binary XOR operation and $\mathrm{bitsum}(\cdot)$ counts the set bits in a bit-vector.

III. The BRIEF-Gist Scene Descriptor

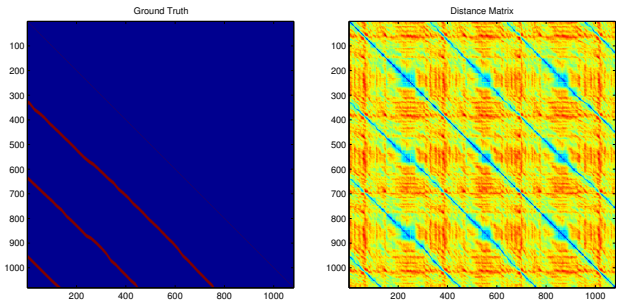
The good recognition performance of BRIEF on local keypoints reported by [3] inspired us to use BRIEF as a holistic descriptor for a complete image. We call this approach *BRIEF-Gist*.

The implementation is very straight-forward: To calculate the BRIEF-Gist descriptor, we first downsample the image to a suitable size close to the descriptor patch size (e.g. 60×60 pixel). Then we calculate the BRIEF descriptor around the center of the downsampled image using OpenCV's [20] implementation.

Another idea is to partition the image in $m \times m$ equally sized tiles. This tiled BRIEF-Gist descriptor is calculated by downsampling the image to a size of $m \cdot s \times m \cdot s$ pixel, where s is the descriptor patch size, e.g. $s = 48$. Then a BRIEF descriptor is calculated for each of the m^2 tiles separately, resulting in m^2 bit-vectors that are stacked to gain the final descriptor vector.

BRIEF-Gist descriptors can be calculated and compared extremely fast: Using a standard desktop PC (Core 2 Duo) and OpenCV 2.2, calculating the 64 bytes long BRIEF-Gist descriptor takes only 1 ms, including the necessary image downsampling and color conversion. The calculation of the BRIEF descriptor itself takes only 0.05 ms. Calculating the similarity between two descriptors according to (1) is performed in 0.001 ms.

The similarity between two scenes, respectively their distance in the descriptor space is given by the distance of their BRIEF-Gist descriptors as defined in (1). Notice that depending on how the scene similarity information is processed further, it can be thresholded to gain a binary decision on whether two scenes are identical. Otherwise the continuous distance value can be used further.



(a) (b)

Fig. 1. (a) Ground truth mask for the New College dataset (scenes $120 \ldots 1200$). Red indicates a manually determined loop closure. (b) Distance matrix built by the individual scene distances δ_{ij} for the same dataset. Notice that the three loop closures are clearly visible as secondary diagonals.

Given the definition of the BRIEF-Gist descriptor, we now want to evaluate how well it performs on the task of scene recognition. The next two sections benchmark the descriptor on two publicly available datasets.

IV. Evaluation – New College Dataset

To quantitatively evaluate the recognition performance of BRIEF-Gist, we first used the recently published *New College Dataset* [23].

This dataset consists of 7854 sets of panoramic images captured by a Ladybug 2 camera, where each of the panoramic images consists of 5 single images, resulting in a total of 39270 single images. We consider each of the 7854 panoramic images a *scene*. The dataset also ships with GPS and odometry data and other information like laser scan messages that are not of importance for our evaluation. The images were collected on a 2.2 km long course over the campus of Oxford New College. The dataset features several loop closings in a dynamic environment including moving people, as well as different types of environment (urban, park) and changing lighting conditions.

A. Ground Truth

Unfortunately, ground truth information is not available. GPS measurements are available for roughly only half of the scenes, but even if GPS is available, it is often disturbed by nearby buildings or vegetation (tree cover etc.). For our first evaluations, we tried to use GPS as ground truth nonetheless. We considered two scenes to be spatially equal if they were closer than a threshold of 7.5 meters as suggested by [12]. However, we found many scenes that were rejected as false positives by the GPS "ground truth" but were manually confirmed to originate from the same spot in the environment by visual inspection. We therefore decided that GPS cannot be trusted as source of ground truth information and thus

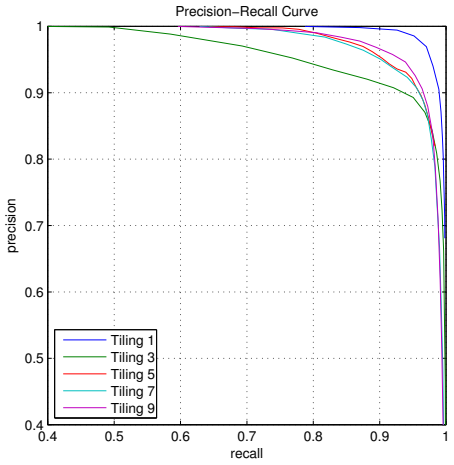


Fig. 2. Precision-recall curve for the New College dataset (scenes 120 . . . 1200) for the BRIEF-Gist descriptor of length 32 and different tilings. Notice the axis scaling.

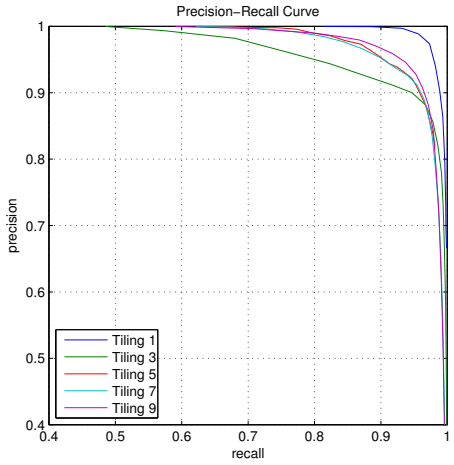


Fig. 3. Precision-recall curve for the New College dataset (scenes 120 . . . 1200) for the BRIEF-Gist descriptor of length 64 and different tilings. Notice the axis scaling. When compared to the results for the descriptor length of 32 bytes (Fig. 2), hardly any differences are visible.

manually determined the best fitting scenes for the first part of the dataset where the robot drove around an enclosure inside the college complex three times (scenes 120 to 1200).

B. Methodology

For scenes 120 to 1200 we manually determined the best matching scene for every 20th scene and linearly interpolated between these fixed matches, assuming constant velocity of the robot.

We calculated the BRIEF-Gist descriptor separately for each of the 5 images associated with every scene s_i, resulting in a set of descriptors $D_{i,k}$ with $i = 120 \ldots 1200$ and $k = 1 \ldots 5$. The distance in appearance space between two scenes s_i and s_j is given by the mean distance of their associated descriptors:

$$\delta_{ij} = \frac{1}{5} \sum_{k=1}^{5} \|D_{i,k} - D_{j,k}\|_H \qquad (2)$$

Here $\|a - b\|_H$ indicates the Hamming distance or L_1 norm as defined in (1). Two scenes s_i and s_j were considered to be equal in appearance space, if their distance δ_{ij} was below a threshold τ. Precision-recall and F-score statistics were generated by varying that threshold τ.

A scene match (s_i, s_j) was considered a true positive if s_j lies within 7 images of the manually determined best match for s_i. That corresponds to a temporal vicinity of approximately 1.5 seconds, as the panoramic images were captured with approximately 5 Hz.

We calculated the precision-recall statistics for varying descriptor lengths (16, 32, and 64 Byte) and tilings (1, 3, 5, 7, 9).

C. Results

Fig. 2 and 3 show the precision-recall plots for descriptor lengths of 32 and 64 bytes respectively. Each plot contains the results for different descriptor tilings. Table I summarizes the recall rates at precisions of 100% and 90%.

TABLE I
RECALL VS. PRECISION, NEW COLLEGE DATASET

Tiling	precision 100%	precision 90%
1	79%	99%
3	48%	93%
5	60%	96%
7	63%	96%
9	60%	97%

The recognition quality on this dataset is surprisingly good. The F-score varies between 0.92 (tiling 3) and 0.97 (tiling 1). It is apparent, that neither the descriptor length, nor the tiling has a very significant influence on the recognition quality.

V. EVALUATION – OXFORD CITY DATASET

We conducted a second quantitative evaluation using the Oxford City Dataset that was published for the Evaluation of FAB-Map [4]. It consists of 1237 image pairs of two cameras facing the forward-left and forward-right of the robot as it was driven through the environment. The cameras captured an image every 1.5 meter.

A. Ground Truth

For this dataset, ground truth information on the loop closures was provided along with the dataset and is depicted in Fig. 5(a).

B. Methodology

Similar to the evaluation on the New College dataset, we calculated the BRIEF-Gist descriptor for the left and right camera image separately. The distance in appearance space between two scenes was formed by the mean of the distances of their respective left and right descriptors.

A scene match was considered a true positive if it was contained in the ground truth matrix. Fig. 5(a) and 5(b) show

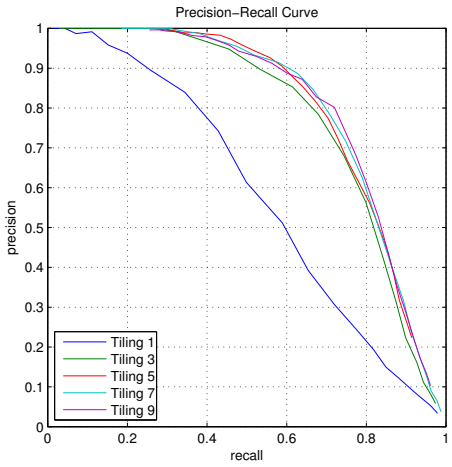
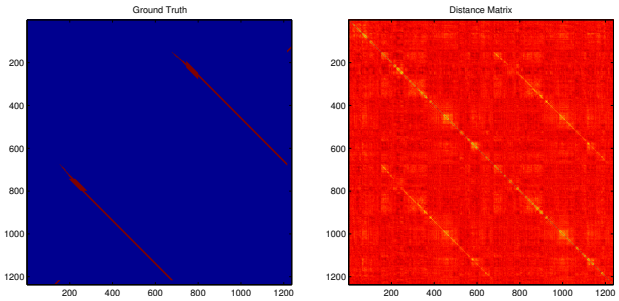


Fig. 4. Precision-recall curve for the Oxford City Centre dataset for the BRIEF-Gist descriptor of length 32 and different tilings.



(a) (b)

Fig. 5. (a) Ground truth matrix for the Oxford City Centre [4] dataset. Red indicates a loop closure between two scenes. (b) Distance matrix built by the individual scene distances δ_{ij} for the same dataset. Notice that the loop closure is clearly visible as secondary diagonal.

the ground truth matrix and the distance matrix calculated by the BRIEF-Gist descriptor.

We calculated the precision-recall statistics for different tilings for the 32 bytes long BRIEF-Gist descriptor.

C. Results

Fig. 4 visualizes the precision-recall statistics. This time, the tiled BRIEF-Gist descriptors are clearly superior to the non-tiled version. For the tiled versions, the F-score varies between 0.72 and 0.75. We found the 7×7 tiling worked best for this dataset, reaching 32% recall at 100% precision. Table II summarizes the recall rates at precisions of 100% and 90% for all tilings.

This performance is comparable to the recall rates of FAB-Map on that dataset. Depending on the exact method chosen, Cummins and Newman reported recall rates of 16%, 31%, or 37% for their system [4] at a precision of 100%. However, the 37% recall was only reached using the most expensive algorithm that takes at least 3 seconds to process a new image. Our results show that BRIEF-Gist is able to perform comparably well, without requiring a dedicated learning step

TABLE II
RECALL VS. PRECISION, OXFORD CITY CENTRE DATASET

Tiling	precision 100%	precision 90%
1	4%	25%
3	26%	53%
5	28%	59%
7	32%	60%
9	23%	58%

to acquire a visual vocabulary and without a computationally involved probabilistic model.

VI. BRIEF-GIST IN A LARGE SCALE SLAM SCENARIO

Encouraged by the evaluation results presented in the previous sections, we wanted to determine if BRIEF-Gist was capable of serving as a front-end in a pose graph SLAM system, solving a large-scale problem.

A. The St. Lucia Dataset

The St. Lucia dataset originally published by Milford et al. [13], consists of 58,758 images taken by a webcam that was mounted on top of a car. The images were collected on a 66 km long course along the roads in St. Lucia, a suburb of Brisbane, Australia. The material (1:40 hours in total) features a dynamic urban environment, changing lighting conditions and many loop closures of different length. Except the video footage, no other sensor information (no odometry, no GPS etc.) is available.

Except the missing ground truth information, we find the dataset is very suitable to evaluate the robustness of BRIEF-Gist. In contrast to the New College dataset, the place recognition system does not have to recognize places that are approached or traversed from different directions. When re-visiting certain streets in the environment, the car always drives in the same direction. Notice that this can be seen as a general weakness of BRIEF-Gist and other appearance-based place recognition systems that use the appearance of the whole scene to perform recognition: They are (in contrast to FAB-Map that relies on distinct landmarks) not invariant to traversal direction.

B. Methodology

We calculated the BRIEF-Gist descriptor with 7×7 tiles for every 5th image of the dataset and calculated the distances between all descriptors. Coarse odometry information was extracted from the images using image profile matching. We used a very similar technique as described in [13], but improved it slightly. Details can be found in [24]. Although this technique is rather simple, the extracted inter-frame motion estimates provide sufficient metric information for the SLAM back-end.

C. A Robust Back-End for SLAM

Given the simplicity of BRIEF-Gist and the results of the evaluation presented before, we cannot guarantee the place recognition to reach a precision of 100%. A certain amount of false positive loop closure detections has to be expected.

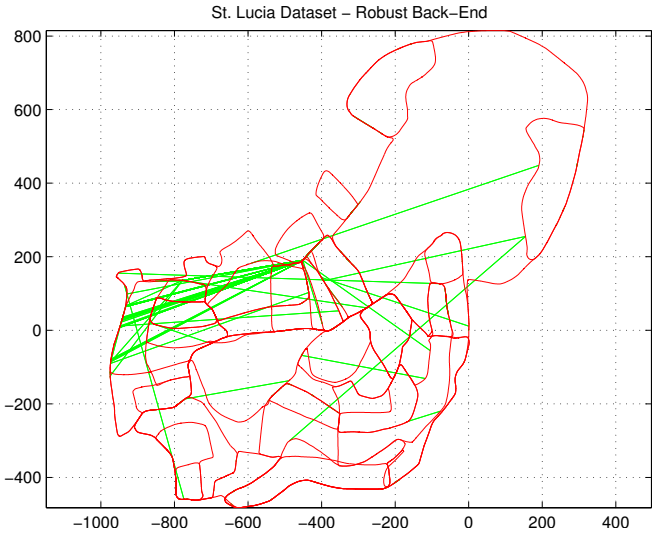


Fig. 6. Final map of St. Lucia. The trajectory is shown in red. Green links indicate false positive loop closures that were erroneously identified by the front-end based on BRIEF-Gist. Our optimization based back-end is robust against erroneous loop closures and correctly rejected them.

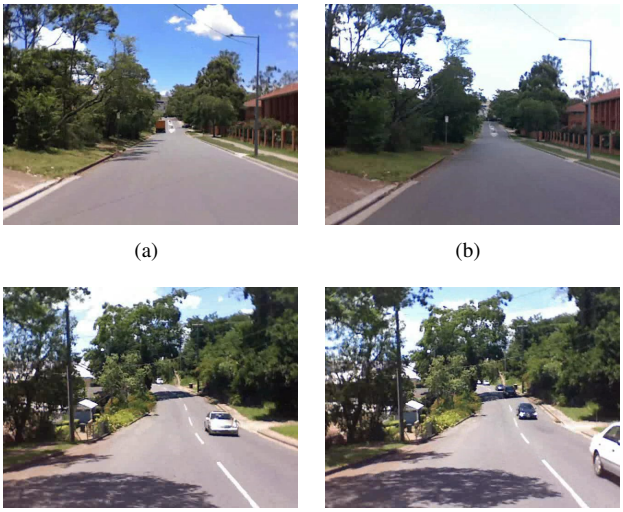


Fig. 7. Examples for correctly matched scenes from the St. Lucia dataset [13]. Despite the significant change in appearance (lighting conditions, moved cars), BRIEF-Gist is able to correctly recognize these scenes and matched the images from (a) with (b), and (c) with (d).

Therefore, the back-end of the SLAM system has to cope with these errors. In parallel work, we developed a robust back-end for pose graph SLAM (based on efficient sparse optimization techniques like [7] or [10]) that is capable of identifying and rejecting these false positive loop closures as part of the optimization process itself. We refer the reader to the appendix for a short introduction to the system.

D. Results

Fig. 6 shows the resulting map after performing SLAM on the whole dataset. As ground truth information are not available, we can only provide a qualitative analysis of the results.

It is apparent that the general structure of the environment has been correctly captured. No false loop closings are present. The front-end based on BRIEF-Gist identified a number of false positive loop closures that were rejected by the back-end during the optimization process. These false positive loop closures are visible as green links in the map of Fig. 6. Some examples of wrongly matched images are shown in Fig. 8.

A small number of loops have not been closed, these are false negative loop closures. In these cases, BRIEF-Gist was not able to recognize the scenes.

The vast majority of the loop closures in the dataset was correctly recognized. This is especially impressive as the scenes over the dataset are visually very similar and ambiguous. Fig. 7 shows a number of exemplary true positive place recognitions.

Fig. 8. Examples for erroneously matched scenes (false positives) from the St. Lucia dataset [13]. BRIEF-Gist incorrectly matched scenes (a) with (b), and (c) with (d) .

VII. Discussion

We presented BRIEF-Gist, a simple scene descriptor based on the BRIEF keypoint descriptor by Calonder et al. [3]. Our evaluation showed that it can compete with state-of-the-art appearance-based place recognition systems like FAB-Map [4]. In contrast to FAB-Map, BRIEF-Gist can be easily implemented, is computationally simple and does not require a learning phase to acquire a vocabulary of visual words. Table III shortly compares BRIEF-Gist and FAB-Map with regard to features and requirements.

BRIEF-Gist is – to a certain extend – invariant to to rotation and displacement, although the BRIEF keypoint descriptor is not. This is because downsampling the input images involves smoothing and interpolation over neighbour-

TABLE III
FEATURE COMPARISON

Invariancy	BRIEF-Gist	FAB-Map [4]
lighting conditions	yes	yes
traversal direction	no	yes
small / large rotations	yes / no	yes / yes
small / large displacement	yes / no	yes / yes
Requirements		
learning phase	no	yes
complex implementation	no	yes
Results		
Oxford City Dataset	recall 32%	recall 16% / 31% / 37%
large-scale SLAM	yes	yes

ing regions in the image. The invariancy is expected to be largest for the non-tiled version of BRIEF-Gist. Invariancy to global changes in the lighting conditions is given because BRIEF is based on a mere comparison of pixel intensity values. The result of these comparisons is not affected by a global change in illumination. We already explained that BRIEF-Gist, like any other place recognition system based on the appearance of the scene as a whole, is not invariant to the direction a scene is traversed. This is in contrast to systems like FAB-Map that work with distinct landmarks. In other words, if the same place is traversed twice, but in different directions, BRIEF-Gist cannot recognize the second encounter as a known place, while FAB-Map can.

We successfully showed that BRIEF-Gist can perform place recognition in the front-end of a pose graph SLAM system in a demanding large-scale SLAM scenario.

In parallel work, we developed a robust optimization-based back-end for pose graph SLAM (see the appendix). Our robust back-end is able to cope with a reasonable number of outliers in the data association, i.e. erroneous loop closure requests. This robustness and the back-end's ability to reject any data association decisions made by the front-end eliminates the need to reach 100% precision (i.e. not a single wrong data association decision) in the place recognition stage. The front-end can therefore be kept simple with regard to computational demands and complexity of the implementation, making BRIEF-Gist a well-suited alternative to more complex systems.

ACKNOWLEDGEMENTS

We thank Michael Milford from Queensland University of Technology for providing the St. Lucia video footage that was presented in his paper [13]. Further material on RatSLAM is available at `http://ratslam.itee.uq.edu.au`.

APPENDIX

In the following we want to shortly introduce our novel robust pose graph SLAM back-end that is based on nonlinear optimization. In contrast to state-of-the-art methods it is, as we have seen above, highly robust against errors in the data association that arise for instance through false positive loop closures.

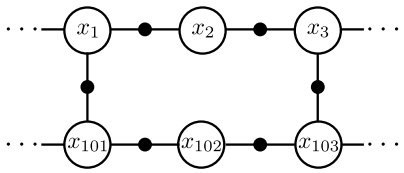


Fig. 9. Factor graph of the standard pose SLAM problem with odometry factors between successive poses and loop closures between (x_1, x_{101}) and (x_3, x_{103}).

A. A General Problem Formulation for Pose Graph SLAM

In the usual probabilistic problem formulation for pose graph SLAM, the goal is to find the optimal (maximum a posteriori) estimate of robot poses X given a set of constraints $U = \{u_{ij}\}$ such that

$$x_j = f(x_i, u_{ij}) + w_{ij} \tag{3}$$

where f is the motion model and w_{ij} is a zero-mean Gaussian with covariance Σ_{ij}. The probability over all variables x_i and constraints u_{ij} is then expressed by

$$P(X, U) \propto \prod_i P(x_j | x_i, u_{ij}) \tag{4}$$

Notice that there are two kinds of constraints u_{ij}: Odometry constraints between successive poses (where $j = i+1$) loop closure constraints u_{ij} that connect non-successive poses and have been determined e.g. by visual place recognition in the front-end.

The optimal estimate on the robot poses, X^*, can be determined by maximizing the joint probability from above:

$$\begin{aligned} X^* &= \arg\max_X P(X, U) \\ &= \arg\min_X -\log P(X, U) \\ &= \arg\min_X \sum_{ij} \|f(x_i, u_{ij}) - x_j\|^2_{\Sigma_{ij}} \end{aligned} \tag{5}$$

Here $\|a - b\|^2_\Sigma$ denotes the squared Mahalanobis distance with covariance Σ, i.e. $\|a - b\|^2_\Sigma = (a - b)^T \Sigma^{-1} (a - b)$.

Given the above formulation, solving (5) and thus finding X^* is left to the back-end. Fig. 9 shows a representation of the problem as a factor graph [9]. Here, the large nodes are the variables x_i and the edges represent the probabilistic constraints (factors) u_{ij} between these variables.

B. Discussion

The above formulation reveals a major problem of current approaches to graph based SLAM. The back-end optimizer has to rely heavily on the front-end to produce a topologically correct factor graph. If the data association step in the front-end fails and erroneously detects a loop closure between two poses x_l and x_k, a factor u_{lk} is introduced between the two corresponding nodes in the factor graph. This factor forces the optimizer to map the two poses onto each other, which will very likely lead to divergence and a defective solution.

A typical strategy to avoid such failures is to apply a sophisticated data association technique. A common approach

based on maximum likelihood and mutual compatibility is the joint compatibility branch and bound algorithm (JCBB) [15]. Olson et al. proposed a compatibility check based on graph partitioning (SCGP) [19], while Cadena et al. use Conditional Random Fields [2]. FAB-Map [4], a probabilistic method for matching scenes in appearance space, is also capable of constructing the loop closure constraints necessary for pose graph SLAM.

However, none of the current data association techniques is guaranteed to work perfectly, i.e. none is guaranteed to reach a precision of 100%. As even a single wrong loop closure constraint can cause the whole SLAM system to fail, the back-end should not have to rely solely on the front-end data association. It should rather be able to change the data association decisions made by the front-end, if they appear to be false at a later time during the optimization.

Our main idea is that the topology of the graph should be subject to the optimization instead of keeping it fixed. If the outlier edges representing data association errors could be identified and removed during the optimization process, the topology could be corrected and thus a correct solution could be reached.

To achieve this, we extend the pose graph SLAM formulation and introduce another kind of variable, the so called *switch factors*. Each of these switches can be understood to act as an variable additional weight on one of the loop closure constraints, with weight values in the interval $(0, 1)$.

The rest of this section explains the details of the implementation.

C. The Robustified Formulation for Pose Graph SLAM

We reformulate (3) and interpret the constraints u_i (with a single index) as control inputs (i.e. odometry readings) between successive poses:

$$x_i = f(x_{i-1}, u_i) + w_i \tag{6}$$

The front end can request a loop closure u_{ij} between two non-successive poses x_i and x_j such that:

$$x_j = f(x_i, u_{ij}) + \lambda_{ij} \tag{7}$$

As above, f is the motion model while w_i and λ_{ij} are zero-mean Gaussian noise terms with covariances Σ_i and Λ_{ij} respectively.

We now introduce a second set of variables $S = \{s_{ij}\}$. Each s_{ij} acts as a switch that controls whether the loop closing constraint u_{ij} (that was proposed by the front-end) between x_i and x_j is accepted by the optimizer or discarded and thus deactivated. Notice that by far not all possible s_{ij} will exist but only those that were proposed by the front-end. The back-end can only deactivate the given loop closure candidates, but never introduce new ones.

Finally, we need to explicitly model a set of switch prior factors $\Gamma = \{\gamma_{ij}\}$ which we explain later on.

Fig. 10 illustrates the general structure of the extended graph.

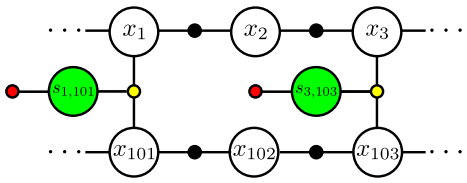


Fig. 10. Factor graph of the proposed extended problem formulation. Depending on the values assigned to the switch variables (s_{ij} shown in green), the loop closure factors (yellow) can be dynamically removed from the graph representation during the optimization process. The switch variables (green) are governed by their prior factors (red) that penalize the deactivation of loop closures.

The joint probability over all variables X, S and measurements U, Γ is given by

$$\begin{aligned} P(X, S, U, \Gamma) \propto & \prod_i P(x_i | x_{i-1}, u_i) \\ & \cdot \prod_{i,j} P(x_j | x_i, u_{ij}, s_{ij}) \\ & \cdot \prod_{i,j} P(s_{ij} | \gamma_{ij}) \end{aligned} \tag{8}$$

We now seek the optimal solution to X and S, namely

$$\begin{aligned} X^*, S^* &= \arg\max_{X,S} P(X, S, U, \Gamma) \\ &= \arg\min_{X,S} -\log P(X, S, U, \Gamma) \\ &= \arg\min_{X,S} \sum_i \| f(x_{i-1}, u_i) - x_i \|^2_{\Sigma_i} \\ &\quad + \sum_{i,j} \| h(x_i, x_j, u_{ij}, s_{ij}) \|^2_{\Lambda_{ij}} \\ &\quad + \sum_{i,j} \| \gamma_{ij} - s_{ij} \|^2_{\Xi_{ij}} \end{aligned} \tag{9}$$

Here f is the motion model as before. The function h and the term γ_{ij} in the above expressions need further explanation.

1) The Loop Closure Factor: We define h as

$$h(x_i, x_j, u_{ij}, s_{ij}) = \text{sig}(s_{ij}) \cdot (f(x_i, u_{ij}) - x_j) \tag{10}$$

where u_{ij} is the spatial displacement between x_i and x_j as indicated by the loop closure detection of the front-end. Furthermore,

$$\text{sig}(a) = \frac{1}{1 + e^{-a}} \tag{11}$$

is the sigmoid function, which implements the desired "switching" behaviour: If s_{ij} indicates an active loop closure then $\text{sig}(s_{ij}) \approx 1$. Thus h penalizes any spatial distance between $f(x_i, u_{ij})$ and x_j and therefore drives the optimizer towards exactly aligning two poses that are connected by a loop closure constraint. By driving s_{ij} towards negative values, the optimizer can "switch off" the loop closure constraint, because in this case $\text{sig}(s_{ij}) \approx 0$ and the spatial distance between $f(x_i, u_{ij})$ and x_j does not add to the global error terms.

The effect of the switch variable can also be understood as acting upon the entries of the information matrix Λ_{ij}^{-1}

that is associated with the loop closure constraint via the Mahalanobis distance $\|h(x_i, x_j, u_{ij}, s_{ij})\|^2_{\Lambda_{ij}}$. Using (10), the definition of the Mahalanobis distance, and the fact that $\mathrm{sig}(s_{ij})$ is a scalar we can write

$$\begin{aligned} \|h(x_i, x_j, u_{ij}, s_{ij})\|^2_{\Lambda_{ij}} &= [\mathrm{sig}(s_{ij}) \cdot A]^T \Lambda_{ij}^{-1} [\mathrm{sig}(s_{ij}) \cdot A] \\ &= A^T \left[\mathrm{sig}(s_{ij})^2 \cdot \Lambda_{ij}^{-1}\right] A \end{aligned} \quad (12)$$

with $A = f(x_i, u_{ij}) - x_j$. In this interpretation, if the variable s_{ij} is driven towards negative values, $\mathrm{sig}(s_{ij}) \approx 0$ and thus the resulting information matrix $\left[\mathrm{sig}(s_{ij})^2 \cdot \Lambda_{ij}^{-1}\right]$ will be close to zero. This however, informally expresses that the constraint A is to be ignored, because literally nothing is known about it, or in other words, the associated uncertainty approaches infinity.

Both interpretations, driving the information measure or the resulting error towards zero, topologically correspond to removing the associated edge from the graph that represents the optimization problem.

2) The Switch Prior Factor: The term γ_{ij} constitutes the prior value of the associated switch factor s_{ij}. In our experiments, we set all $\gamma_{ij} = 10$, as $\mathrm{sig}(10) \approx 1$. This means that we initially accept all loop closure constraints that were proposed by the front-end. During the optimization the Mahalanobis distance $\|s_{ij} - \gamma_{ij}\|^2_{\Xi_{ij}}$ in (9) penalizes the deviation of s_{ij} from its initial value γ_{ij} and thus penalizes the deactivation of a loop closure.

D. Implementation and Parameters

We implemented our approach to a robust back-end in C++ using the GTSAM framework that is available upon request from the group of Frank Dellaert[1]. There is only one free parameter that needs to be set: The covariance matrix Ξ is used in the similarity constraint $\|s_{ij} - l_{ij}\|^2_{\Xi_{ij}}$ in (9): It is a one-dimensional variance measure and was empirically set to $\Xi = 20^2$ for all experiments described in this paper. Ξ controls the penalty for switching off a loop closure during the optimization. The other covariance matrices Λ and Σ are used to calculate the Mahalanobis distances in the odometry and loop closure factors and have to be provided by the front-end.

All experiments were conducted in an incremental fashion, i.e. data was fed into the optimizer 200 frames at a time, in contrast to performing batch optimization.

E. Conclusions

Our modified problem formulation can be understood as transferring parts of the responsibility for correct data association from the front-end into the back-end. The back-end optimizer can now change the topological structure of the pose graph representation during the optimization process. Therefore, it can account for possible data association errors and ignore erroneous loop closure constraints. As the overall SLAM system becomes tolerant and robust against errors in the data association, a reasonable false positive rate is acceptable (precision $< 100\%$) and the data association algorithm and can be kept comparably simple.

[1] `https://collab.cc.gatech.edu/borg/gtsam/`

References

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Proceedings of the ninth European Conference on Computer Vision*, May 2006.

[2] César Cadena, Dorian Gálvez-López, Fabio Ramos, Juan D. Tardós, and José Neira. Robust Place Recognition with Stereo Cameras. In *IEEE Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2010.

[3] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary Robust Independent Elementary Features. In *European Conference on Computer Vision (ECCV)*. Springer, 2010.

[4] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[5] Mark Cummins and Paul Newman. Highly Scalable Appearance-Only SLAM – FAB-MAP 2.0. In *Robotics Science and Systems*, 2009.

[6] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert. iSAM2: Incremental smoothing and mapping with fluid relinearization and incremental variable reordering. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, 2011.

[7] M. Kaess, A. Ranganathan, and F. Dellaert. iSAM: Incremental Smoothing and Mapping. *IEEE Transactions on Robotics*, 24(6), 2008.

[8] K. Konolige, G. Grisetti, R. Kümmerle, W. Burgard, B. Limketkai, and R. Vincent. Efficient sparse pose adjustment for 2d mapping. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2010.

[9] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, February 2001.

[10] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. g2o: A general framework for graph optimization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2011.

[11] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. In *International Journal of Computer Vision, 60, 2*, 2004.

[12] Will Maddern, Michael Milford, and Gordon Wyeth. Continuous Appearance-based Trajectory SLAM. In *International Conference on Robotics and Automation (ICRA)*, 2011.

[13] Micheal J. Milford and Gordon F. Wyeth. Mapping a Suburb with a Single Camera using a Biologically Inspired SLAM System. *IEEE Transactions on Robotics*, 24(5), October 2008.

[14] A.C. Murillo and J. Kosecka. Experiments in place recognition using gist panoramas. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2196 –2203, 2009.

[15] J. Neira and J.D. Tardos. Data association in stochastic mapping using the joint compatibility test. *IEEE Transactions on Robotics and Automation*, 17(6):890–897, 2001.

[16] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the IIEEE Conference on Computer Vision and Pattern Recognition*, pages 2161–2168. IEEE Computer Society, 2006.

[17] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Visual Perception, Progress in Brain Research*, 155, 2006.

[18] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3), 2001.

[19] Edwin Olson, Matthew Walter, Seth Teller, and John Leonard. Single-cluster spectral graph partitioning for robotics applications. In *Robotics: Science and Systems (RSS)*, 2005.

[20] OpenCV. The OpenCV Library. `http://opencvlibrary.sf.net`.

[21] Grant Schindler, Matthew Brown, and Richard Szeliski. City-scale location recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–7, 2007.

[22] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, 2003.

[23] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The New College Vision and Laser Data Set. *International Journal for Robotics Research (IJRR)*, 28(5):595–599, May 2009.

[24] Niko Sünderhauf and Peter Protzel. Beyond RatSLAM: Improvements to a Biologically Inspired SLAM System. In *Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation*, 2010.