Early Bird: Loop Closures from Opposing Viewpoints for Perceptually-aliased Indoor Environments

Satyajit Tourani[1], Dhagash Desai[1], Udit Singh Parihar[1], Sourav Garg[2],
Ravi Kiran Sarvadevabhatla[3], Michael Milford[2] and K. Madhava Krishna[1]
[1]*Robotics Research Center, IIIT Hyderabad, India*
[2]*Centre for Robotics, Queensland University of Technology (QUT), Australia*
[3]*Centre for Visual Information Technology, IIIT Hyderabad, India*

Keywords: Visual Place Recognition, Homography, Image Representation, Pose Graph Optimization, Correspondences Detection.

Abstract: Significant recent advances have been made in Visual Place Recognition (VPR), feature correspondence and localization due to deep-learning-based methods. However, existing approaches tend to address, partially or fully, only one of two key challenges: viewpoint change and perceptual aliasing. In this paper, we present novel research that simultaneously addresses both challenges by combining deep-learnt features with geometric transformations based on domain knowledge about navigation on a ground-plane, without specialized hardware (e.g. downwards facing cameras, etc.). In particular, our integration of VPR with SLAM by leveraging the robustness of deep-learnt features and our homography-based extreme viewpoint invariance significantly boosts the performance of VPR, feature correspondence and pose graph sub-modules of the SLAM pipeline. We demonstrate a localization system capable of state-of-the-art performance despite perceptual aliasing and extreme 180-degree-rotated viewpoint change in a range of real-world and simulated experiments. Our system is able to achieve early loop closures that prevent significant drifts in SLAM trajectories.

1 INTRODUCTION

Visual Place Recognition (VPR) and local feature matching are an integral part of a visual SLAM system for correcting the drift in robot's trajectory via loop closures. However, multiple complicating factors make this process challenging such as variations in lighting and viewpoint along with the need to deal with dynamic objects.

Typically, indoor structures (e.g. walls, ceilings) tend to be feature-deficient. They often exhibit strong self-similarity, leading to perceptual aliasing. Ergo, VPR becomes further challenging when a place is revisited from a very different viewpoint eg. an opposing viewpoint (180° viewpoint shift). The latter is a situation commonly encountered when tackling VPR for indoor-based scenarios in warehouses, office buildings and their corridors.

Due to the above mentioned challenges, existing state-of-the-art place representation methods struggle to perform well. In particular, deep learning-enabled viewpoint-invariant global image representations (Arandjelovic et al., 2016; Garg et al., 2018b) are unable to deal with perceptual aliasing due to repetitive indoor structures. Whereas, viewpoint-presumed image representations (Dalal and Triggs, 2005) that retain spatial layout of the image fail due to 180° viewpoint shift, as also demonstrated in (Garg et al., 2018a). Therefore, a robust place representation leveraging discriminative regions of an image is much needed to deal with this problem.

While seemingly aliased, in practice, floor patterns contain discriminative features. Blemishes, scratches on the floor surface and natural variations in floor/ground surfaces yield features which can be detected easily across conditional variations (Zhang et al., 2019). In turn, these enable reliable localization (Kelly et al., 2007; Nourani-Vatani et al., 2009). However, most of the existing solutions based on floor patches require specialised hardware (e.g. downward-facing cameras (Nourani-Vatani et al., 2009; Mount and Milford, 2017), additional light sources (Kelly et al., 2007)).

Tourani, S., Desai, D., Parihar, U., Garg, S., Sarvadevabhatla, R., Milford, M. and Krishna, K.
Early Bird: Loop Closures from Opposing Viewpoints for Perceptually-aliased Indoor Environments.
DOI: 10.5220/0010230804090416
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 409-416
ISBN: 978-989-758-488-6

We propose an indoor VPR approach which addresses the concerns highlighted previously. With our proposed pipeline, we make the following contributions:

- A novel pipeline that combines projective geometry and deep-learnt features to focus specifically on floor areas and consistently improves the following pipeline modules: VPR (able to deal with opposing viewpoints), feature correspondences; leading to improved inputs for subsequent SLAM pose-graph optimization.
- Extensive comparisons across various deep architectures showing that VPR and feature correspondence modules suffer significantly when used on raw images while achieving a significant boost in performance when used on rotationally-aligned floor areas. The improvement is consistent over various real and simulated floor types as shown in section 5.
- The paper unveils Early Bird SLAM that integrates the above VPR and feature-correspondence pipeline in a back-end pose graph optimizer, demonstrating substantial decrease in Absolute Trajectory Error (ATE) as compared to the state-of-the-art SLAM frameworks such as (Labbé and Michaud, 2013).

2 RELATED WORK

2.1 Descriptor based Recognition

Amongst earlier methods, the most popular were appearance-based place descriptors such as Bag of Visual Words (BoVW) (Sivic and Zisserman, 2003; Csurka et al., 2004) and Vector of Locally Aggregated Descriptors (VLAD) (Jégou et al., 2010) where a visual vocabulary is constructed using local features like SURF (Bay et al., 2008) and SIFT (Lowe, 1999). These have been used in FAB-MAP (Cummins and Newman, 2008) and ORB-SLAM (Mur-Artal et al., 2015) to good effect.

Whole-image descriptors like Gist (Oliva and Torralba, 2006) and HoG (Dalal and Triggs, 2005) presume the scene viewpoint to remain similar across subsequent visits of the environment, enabling VPR under extreme appearance variations as demonstrated in SeqSLAM (Milford and Wyeth, 2012).

2.2 Robustifying VPR

Many solutions have been proposed to robustify VPR to viewpoint variation.

CNNs with their partial viewpoint invariance have been shown to robustify VPR ((Sünderhauf et al., 2015)). They allow for end-to-end training where one can in addition to using off-the-shelf networks (Arandjelovic et al., 2016), train the later layers to obtain task/dataset specific results (Radenović et al., 2018). In (Chen et al., 2017), pyramid pooling was shown to improve viewpoint robustness.

Opposing Viewpoints. Most of the existing literature that addresses viewpoint-invariance for VPR assumes a large amount of visual overlap.

LoST (Garg et al., 2018b) used dense semantic information to represent places and extract keypoints from within the CNN to enable high-performance VPR. This was improved upon in (Garg et al., 2019) using a topo-metric representation of places. In the vein of utilizing higher-order semantics, X-view (Gawel et al., 2018) uses dense semantic segmentation and graph-based random walks to perform VPR.

Saliency of Floor Features. Floor features have been shown to be salient enough to aid in VPR. In (Zhang et al., 2019), features are extracted from floor surfaces to perform global localization. The imperfections in the tiles provide enough features that keypoints can be extracted.In (Mount and Milford, 2017) the authors have explored surface based localization methods for match verification using ground-based imagery. (Kelly et al., 2007) proposed to use floor patches to perform local region matching in order to develop an infrastructure-free localization system. In (Nourani-Vatani et al., 2009), authors developed a visual odometry system based on floor patches.

Keypoint Correspondences. Classical approaches like SIFT (Lowe, 2004) and SURF (Bay et al., 2008) tackle the problem of calculating the pixel level correspondences between images in a two-way approach by first detecting the keypoints and then describing a local region around the keypoint. Recent learning-based approaches like SuperPoint (DeTone et al., 2018) and D2Net (Dusmanu et al., 2019) combine detection and description by simultaneously optimizing for both the tasks. However, none of the approaches work well when deployed in perceptually-aliased and low-textured indoor settings, particularly when viewing a scene from an opposite direction. We show that the existing deep-learnt feature correspondence methods can lead to better matching by using certain regions of the image like floor and exploiting geometric priors between images in the forward and reverse trajectory.

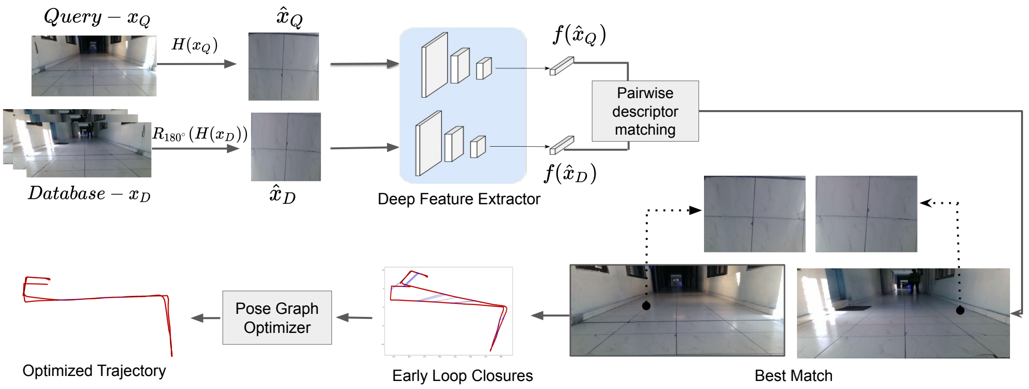

Figure 1: Our proposed pipeline for Visual Place Recognition. Images are first converted to a top view by homography. This transformed view is fed into a deep feature extractor to obtain the output descriptor which forms our place representation. Then, a cosine-distance based cost matrix is constructed to find matches. A matched pair is then fed into a feature correspondence extractor to find correspondences. This is subsequently fed into a pose graph optimizer to obtain optimized trajectory.

3 METHODOLOGY

Our proposed hierarchical pipeline consists of the following three stages: indoor visual place recognition for opposite viewpoints, feature correspondence extraction and pose graph optimization.

3.1 Indoor Visual Place Recognition for Opposite Viewpoints

While the method employs deep-learned features, it requires no training of the underlying feature extractor.

Our approach to indoor VPR utilises the fact that floor patches contain useful features in the form of cracks, designs, dirt/stains. The floor-based features act as a unique signature for specific places within an indoor region.

To extract the floor-region of the images taken, we fit a planar homography H to image points via a RANSAC + 4 point algorithm (Hartley and Zisserman, 2003). We use a fixed homography matrix across all the datasets. In the original image, we pick four points along the floor region which are then transformed into a floor image.

Let H be the homography matrix, x be a homogenized coordinate of an input image then the transformed image co-ordinate, $\hat{x}$ is obtained via eq. (1).

$$\hat{x} = H(x) \tag{1}$$

Figure 2 shows example images from our benchmark datasets with both the raw images and their corresponding floor patches so obtained.

In our pipeline, we pass the floor patch images obtained into a deep feature extractor and the output descriptor forms our place representation.

$$d_i^Q = f(\hat{x}_i^Q) \tag{2}$$

where $f()$ corresponds to the process of obtaining features from a deep feature extractor and d_i^Q is the resultant descriptor obtained for image i. Cosine distance-based descriptor matching is done to obtain matches between Q and D. We apply the homography operation on the reference and then perform a 180° rotation of the transformed images to improve matching across differing viewpoints of the same place. Although the rotation/flipping operation is not necessary for some of the deep feature extractors as they are inherently viewpoint-invariant, we show that performance can be boosted for such descriptors whereas other viewpoint-presumed deep feature description techniques become only useful post image rotation.

3.2 Feature Correspondence

By utilizing the previously proposed concept of applying geometric transformations on image to extract textured floor regions enables us to generate very precise pixel level correspondences Figure 4 (1c and 2c). These precise correspondences are also very essential to calculate near ground truth transformation and subsequent loop closure in pose graph SLAM Figure 5 on real dataset and Figure 6 on synthetic dataset.

Let x_Q be the query image and x_M be the matched image from the opposite trajectory obtained via the VPR pipeline. $\hat{x}_Q$ is the transformed image obtained by applying homography and $\hat{x}_M$ is the transformed image obtained by applying homography and

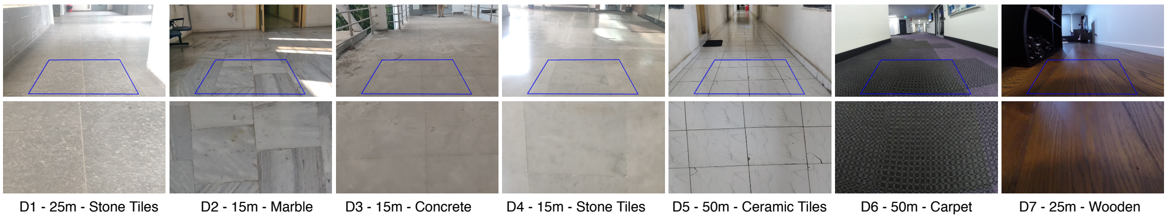

Figure 2: Example images from our indoor dataset. We demonstrate the usage of our indoor VPR pipeline on various floor types.(First row) Raw images. (Second row) Homography transformed images of the respective raw images in the first row. Below we also mention the dataset id along with its trajectory length in meters and floor type in the form of (ID-Length-Floor Type). We overcome challenges such as self repetition and lack of features by using homography transformed floor images which contain distinctive features that are helpful in performing VPR.

π-rotation, Figure 3. Local feature extractor $g(.)$, in our case D2Net is used to obtain correspondences $\hat{q}^{2D}$ and $\hat{m}^{2D}$ on transformed images. Correspondences on the original image q^{2D} and m^{2D} are obtained by inverse π-rotation and inverse homography.

$$\hat{x}_Q = H(x_Q) \tag{3}$$

$$\hat{x}_M = R_\pi(H(x_M)) \tag{4}$$

$$\hat{q}^{2D}, \hat{m}^{2D} = g(\hat{x}_Q, \hat{x}_M) \tag{5}$$

$$q^{2D} = H^{-1}(\hat{q}^{2D}) \tag{6}$$

$$m^{2D} = H^{-1}(R_\pi^{-1}(\hat{m}^{2D})) \tag{7}$$

3.3 Pose Graph Optimization

The proposed VPR pipeline has direct applicability in loop closure or data association problem in visual SLAM. Formally, we are interested in finding the optimal configuration X^* of robot poses x_i based on odometry constraints u_i and loop closing constraints c_{qm}. Here, odometry constraints u_i are used to build the motion model whereas loop closure constraints c_{qm} provide information to correct the error accumulated due to sensors' noise.

Let S be a set of image pairs proposed by VPR such that, $S = \{(q,m)|I_q \in Q, I_m \in R\}$, then optimal poses $\mathtt{X}^*$ are given by:

$$X^* = \underset{X}{argmax}\, P(X|U,C) = \underset{X}{argmax} \underbrace{\prod_i P(x_{i+1}|x_i,u_i)}_{Odometry\ Constraints} \times \underbrace{\prod_{(q,m)\in S} P(x_m|x_q,c_{qm})}_{Loop\ Closure\ by\ VPR} \tag{8}$$

2D correspondences q^{2D} and m^{2D} obtained using D2Net in the previous subsection, are projected into $3D$ using the camera matrix K and depth λ, and finally $3D$ points are registered using ICP. The ICP recovered transform c_{qm} between the two images form the loop closure constraint in the pose graph optimizer.

$$Q^{3D} = \lambda K^{-1} q^{2D} \tag{9}$$

$$M^{3D} = \lambda K^{-1} m^{2D} \tag{10}$$

$$c_{qm} =< R, T >= ICP(Q^{3D}, M^{3D}) \tag{11}$$

To ensure that only high-precision loop closure constraints are used in pose graph optimization, we shortlist query-reference image pairs based on their cosine distance. We used only top 20 loop closure pairs with lowest cosine distance. Cauchy robust kernel in cost function is used to minimize the effect of false positive loop closures that might have crept in pose graph, using pose graph optimizer g2o (Kümmerle et al., 2011).

4 EXPERIMENTAL SETUP

Datasets. We have collected seven real-world indoor datasets in our experiments as shown in Figure 2. Six in a university campus and one inside a home. The datasets comprise different types of floor types like marble, wooden, concrete and carpet. The datasets consist of sequences in range of 15 m to 50 m. Each sequence contains anywhere between 500-4000 images. Five were collected using a OnePlus 6 and two were collected using GoPro Hero 3+. We have shown the performance of our VPR pipeline in a SLAM framework, Figure 5, on one of the dataset collected on the university campus with P3DX robot equipped with RealSense D435 and wheel odometer. Ablation studies of the effect of loop closures on the ATE of SLAM pose graph are done on three synthetic datasets, Figure 6 (last column), where floor tiles are chosen from real world images and P3DX noise model have been incorporated in the simulator odometry data.

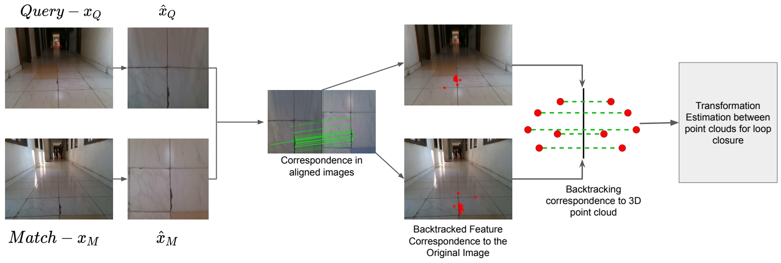
$Query - x_Q$
$\hat{x}_Q$
$Match - x_M$
$\hat{x}_M$
Correspondence in aligned images
Backtracked Feature Correspondence to the Original Image
Backtracking correspondence to 3D point cloud
Transformation Estimation between point clouds for loop closure

Figure 3: Pipeline depicting the process of transformation estimation from proposed opposite pairs by VPR.

4.1 Evaluation and Comparisons

We use Recall as an evaluation metric for visual place recognition, defined as the ratio of true positives and total number of positives. A match is said to be a true positive if it lies within a localization radius of $\frac{1}{15}$th of the total length of the traversal of its ground truth. We compare various deep feature extractors under different input settings.

For the feature correspondence results, performance of different types of input transformations is shown qualitatively and quantitatively in Figure 4 and Table 3 respectively. A correspondence match is considered to be an inlier if its reprojection error calculated using ground truth transformation is below a threshold. Comparison among classical approach SURF (Bay et al., 2008) and learning based approaches SuperPoint (DeTone et al., 2018) and D2Net (Dusmanu et al., 2019) has been shown in Table 3.

We have compared *Early Bird* loop closures with state of the art SLAM method RTABMAP (Labbé and Michaud, 2013) in terms of Absolute Trajectory Error(ATE) on both real world and synthetic datasets.

Table 1: Quantitative Analysis: First column shows the preprocessing applied to the input. Second column shows types of transformation applied on the reference images. π-Rot and Flip-LR indicate a 180° rotation and horizontal left-to-right flipping of the reference image respectively. Homo + π-Rot gives the best results in most cases. NetVLAD is used as the deep feature extractor. D1-D7 are the datasets as mentioned in Figure 2.

Input	OP	D1	D2	D3	D4	D5	D6	D7
Raw	None	24.1	22.1	24.8	19.9	9.7	28.5	22.2
Raw	Flip-L-R	28.8	26.8	29.1	26.6	12.5	**29.2**	20.6
Homo	None	60.7	62.7	61.3	70.4	40.1	12.2	15.1
Homo	π-Rot	**69.3**	**70.1**	**71.8**	**73.3**	**44.7**	16.3	**28.8**

5 RESULTS

We show results for each of the components of our pipeline, particularly highlighting the effect of geometric transformations for both VPR and feature correspondences which ultimately contribute in improving the trajectory error for the SLAM back end. First, we show results for VPR with ablations across many descriptors and geometric transformations followed by qualitative and quantitative results for feature correspondence extraction. Finally, we compare our Early Bird SLAM pipeline with the state-of-the-art SLAM system RTABMAP in terms of Absolute Trajectory Error (ATE) on real and synthetic datasets.

5.1 Visual Place Recognition

Table 1 and 2 show the recall performance for VPR using seven different datasets. While Table 1 highlights the effect of geometric transformations on a given place descriptors, NetVLAD in this case, Table 2 compares different descriptor types for the best performing geometric transformation, that is, *Homo* + *π-Rot*. It can be observed in Table 1 that using the raw images (*Raw*) as input leads to inferior results for most of the datasets even when using the state-of-the-art viewpoint-invariant representation NetVLAD. The best performance is achieved achieved through 180-degree rotation of floor patches (*Homo* + *π-Rot*) as compared to when using only the homography transformed input (*Homo*). We also compute descriptors using horizontally-flipped images (Flip L-R) as used in (Garg et al., 2018a; Garg et al., 2018b) for dealing with opposing viewpoints in outdoor environments. It can be observed that such a transformation does not lead to consistent performance gains. We attribute this to the repetitive and featureless nature of indoor environments. For *D7*, a small performance gap is observed between using *Raw* and geometrically transformed images (*Homo* + *π-Rot*); this is due to the reduced aliasing because of availability of unique visual landmarks when using raw images. In *D6*, raw images (*Raw*) perform better than floor (*Homo* + *π-Rot*) images due to the lack of sufficient visual features on the carpet floor. This limitation could potentially be overcome by using a joint VLAD aggregation of the whole image and the transformed image, and remains a future work.

Table 2: Quantitative Analysis: We compare the performance of various deep feature extractors where planar homography and 180° rotation (Homo+π-rot) is applied to the database images. D1-D7 are the datasets as mentioned in Figure 2.

Models	D1	D2	D3	D4	D5	D6	D7
NetVLAD	69.3	**70.1**	**71.8**	**73.3**	**44.7**	16.3	**28.8**
Resnet	**72.8**	66.5	60.6	63.3	38.0	16.9	13.2
VGG-19	71.6	63.0	56.0	61.1	32.5	**17.6**	5.0
Superpoint	16.5	27.6	15.0	23.0	35.7	1.6	3.2

In Table 2 we compare the performance of different feature extractors under the best input setting (*Homo* + *π-Rot*). NetVLAD performs the best in most cases with ResNet being the second best. While NetVLAD is a viewpoint-invariant representation, ResNet-based feature extraction assumes the viewpoint to be the same after geometric transformations. A viewpoint-invariant representation has more advantages which is reflected in Table 2. Nevertheless, due to high perceptual aliasing, geometric transformations are *still* required before descriptor computation in order to achieve the best performance, as demonstrated in Table 1.

5.2 Feature Correspondence

Estimating precise correspondences are crucial to calculate accurate transformations using ICP like registration methods, which in turn help us to achieve near ground truth pose estimates. Figure 4 (1a and 2a) show that using raw images to calculate feature correspondences cause both SURF and state-of-the-art learning methods SuperPoint and D2Net to fail. The number of correct correspondences increase with the use of geometric transformations focusing on textured floor regions. However, without image rotation, matching still remains poor as shown in Figure 4 (1b and 2b). The best results are obtained when transformed image pair is aligned with each other by 180° rotation as shown in Figure 4 (1c and 2c). Table 3 quantitatively shows the number of inliers as well as total correspondences, averaged over all the datasets, using different feature extractor methods. It can be observed that *after* the geometric transformation, D2Net leads to a large number of initial as well as final correspondences. Thus, we used D2Net with homography and π-rotation as the final keypoint extractor approach for calculating transformations in subsequent tasks.

5.3 Loop Closures in Early Bird SLAM

In practical scenarios in the context of long term autonomy, a robot can typically revisit its operating environment from a variety of different viewpoints. A more common scenario particularly in corridors and aisles is that of an opposite viewpoint which is when our proposed system triggers loop closure. We demonstrate its efficacy by comparing with the state-of-the-art SLAM system RTAB-Map (Labbé and Michaud, 2013). As shown in the Table 4 (Dataset D5), we significantly reduce the Average Trajectory Error (ATE) by detecting "early"

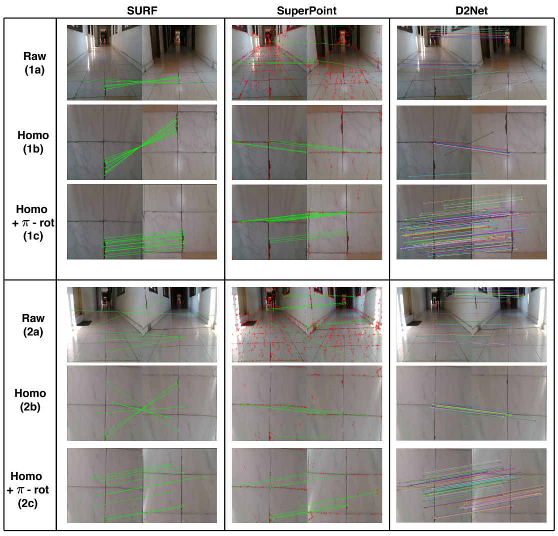

Figure 4: Comparison of correspondences obtained using raw image vs homo + π-rot. The best set of correspondences is obtained using D2Net with homo + π-rot operation.

Table 3: Number of Inliers / Total Correspondences averaged over all datasets.

	SURF	Superpoint	D2Net
Raw	4.5 / 6.5	0 / 6.5	0 / 28.5
Homo	8.5 / 9	4 / 5.5	6.5 / 13
Homo+π-Rot	**11.5 / 11.5**	**10 / 11.5**	**97.5 / 97.5**

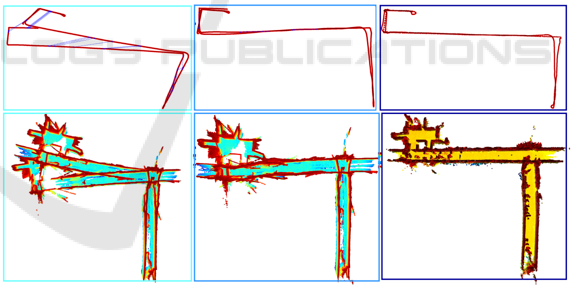

Figure 5: Rows represents pose graph and registered map.The first column corresponds to RTABMAP trajectory with robot revisiting the location from opposite viewpoints, blue dashed line represents early loop closures on pose graph, second and third column corresponds to optimized map based on VPR constraints and ground truth map respectively.

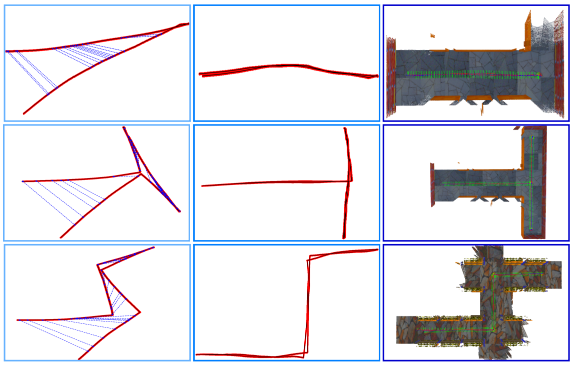

Figure 6: Each row represents pose graph optimization done on three gazebo environments corresponding to Table 4. First column corresponds to RTABMAP trajectory with robot revisiting the location from opposite viewpoints. Second and third column corresponds to optimized and ground truth trajectory.

loop closures which is qualitatively shown in Figure 5 while RTABMAP fails to do so, due to significant perceptual aliasing. As a consequence, the RTABMAP's trajectory Figure 5 (*top left*) shows multiple corridors when actually there is only one. The *(middle)* column Figure 5 shows trajectory and top view map due to loop detection and closure from opposite views using our VPR pipeline. These are much closer to ground truth trajectory and map shown in *(last)* column of Figure 5. Ablation study has been done on measuring the usefulness of early loop closures with different robot's trajectory length and different tile patterns. We have performed our experiment in three different simulated environment settings as shown in Figure 6. In all the three cases, we have achieved near ground truth poses, while state of the art SLAM method RTABMAP fails to detect *any* loop closure while returning to the same place from opposite viewpoints. These results are quantitatively represented in terms of ATE in Table 4, showing that with increase in the length and complexity of the trajectory the effect of *early* loop closures becomes even more pronounced.

Table 4: Average Trajectory Error on a university dataset and three gazebo datasets for RTABMAP loop closures vs Early Bird loop closures.

Datasets	RTABMAP	Early Bird
D5	9.239	**5.69**
S1	0.94	**0.26**
S2	2.86	**0.38**
S3	2.85	**0.36**

6 CONCLUSION AND FUTURE WORK

This paper proposes a novel pipeline that integrates Visual Place Recognition (VPR) with SLAM front and back ends specifically for loop detection and closure for opposite views. The paper showcases that rotationally-aligned deep floor descriptors provide for significant boost in all the submodules of the pipeline: VPR (loop detection), descriptor matching and pose graph optimization.

The paper extensively compares several deep architectures for VPR and descriptor matching. Given the geometric transformations, NetVLAD as a global descriptor was found most suitable for opposite-view VPR while descriptor matching through D2Net found maximal number of matches vis a vis competing descriptor matching frameworks. Also, superior place recognition and descriptor matching across opposite views resulted in a similar performance gain in back-end pose graph optimization. Specifically, we showed *early* loop closures that prevented significant drifts in SLAM trajectories as a consequence of the proposed pipeline along with the proper choice of deep architectures that exploit the various sub-modules of the pipeline to its maximum efficacy.

The future threads include extension to outdoor and warehouse like topologies, use of visual semantics or monocular depth-based ground plane extraction and learning an attention mechanism to deal with the simultaneous effect of viewpoint variations and perceptual aliasing.

REFERENCES

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359.

Chen, Z., Jacobson, A., Sunderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I., and Milford, M. (2017). Deep learning features at scale for visual place recognition. pages 3223–3230.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.

Cummins, M. and Newman, P. (2008). Fab-map: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

DeTone, D., Malisiewicz, T., and Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. pages 337–33712.

Dusmanu, M., Rocco, I., Pajdla, T., Pollefeys, M., Sivic, J., Torii, A., and Sattler, T. (2019). D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Garg, S., Babu, M., Dharmasiri, T., Hausler, S., Suenderhauf, N., Kumar, S., Drummond, T., and Milford, M. (2019). Look no deeper: Recognizing places from opposing viewpoints under varying scene appearance using single-view depth estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4916–4923. IEEE.

Garg, S., Suenderhauf, N., and Milford, M. (2018a). Don't look back: Robustifying place categorization for viewpoint-and condition-invariant place recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3645–3652. IEEE.

Garg, S., Suenderhauf, N., and Milford, M. (2018b). Lost? appearance-invariant place recognition for opposite viewpoints using visual semantics. *arXiv preprint arXiv:1804.05526*.

Gawel, A., Del Don, C., Siegwart, R., Nieto, J., and Cadena, C. (2018). X-view: Graph-based semantic multi-view localization. *IEEE Robotics and Automation Letters*, 3(3):1687–1694.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

Jégou, H., Douze, M., Schmid, C., and Pérez, P. (2010). Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE.

Kelly, A., Nagy, B., Stager, D., and Unnikrishnan, R. (2007). Field and service applications-an infrastructure-free automated guided vehicle based on computer vision-an effort to make an industrial robot vehicle that can operate without supporting infrastructure. *IEEE Robotics & Automation Magazine*, 14(3):24–34.

Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., and Burgard, W. (2011). g 2 o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pages 3607–3613. IEEE.

Labbé, M. and Michaud, F. (2013). Appearance-based loop closure detection for online large-scale and long-term operation. *IEEE Transactions on Robotics*, 29(3):734–745.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110.

Milford, M. J. and Wyeth, G. F. (2012). Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation*, pages 1643–1649. IEEE.

Mount, J. and Milford, M. (2017). Image rejection and match verification to improve surface-based localization. In *Proc. Australas. Conf. Robot. Autom.*, pages 213–222.

Mur-Artal, R., Montiel, J. M. M., and Tardos, J. D. (2015). Orb-slam: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, 31(5):1147–1163.

Nourani-Vatani, N., Roberts, J., and Srinivasan, M. V. (2009). Practical visual odometry for car-like vehicles. In *2009 IEEE International Conference on Robotics and Automation*, pages 3551–3557. IEEE.

Oliva, A. and Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36.

Radenović, F., Tolias, G., and Chum, O. (2018). Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of International Conference on Computer Vision (ICCV)*, page 1470. IEEE.

Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., and Milford, M. (2015). Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. *Proceedings of Robotics: Science and Systems XII*.

Zhang, L., Finkelstein, A., and Rusinkiewicz, S. (2019). High-precision localization using ground texture. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6381–6387. IEEE.