Supplementary Material: Binary Graph Descriptor for Robust Relocalization on Heterogeneous Data

Xi Wang, Marc Christie, Eric Marchand

Supplementary Material: Binary Graph Descriptor for Robust Relocalization on Heterogeneous Data

Xi Wang, Marc Christie, Eric Marchand

I. Geometrical Accuracy in Comparison with Local Features

In this section, we discuss the geometrical accuracy of the proposed BIG descriptor and compare the reprojection error and rotational error with other local features *e.g.,*. ORB [1], SIFT [2], SURF [3].

Two datasets are evaluated in this section: i) Newer College Dataset [4], a dataset of grey-level and Lidar information, recorded in well-lit and consistent condition; ii) RobotCar seasonal Dataset [5], provides car-mounted RGB images, recorded under different seasonal, weather conditions and also influenced by the day-night shift. We applied image + NetVLAD [6] and FGSN [7] semantic image layers for generating BIG descriptor in two datasets respectively.

We take all correspondent loop-closures (*i.e.* image pairs) from ground-truth and estimate their Fundamental matrix by extracting and matching features from each pair of images, with the help of the RANSAC method in OpenCV. The reprojected error is represented in pixel by computing a mean and a median on all matched features' error, instead of using inliers features to avoid overfitting the RANSAC technique. Decomposed rotational errors (in degree) are reported as well for evaluating the geometrical accuracy of the proposed binary graph feature; the reason for ignoring the translational vector is due to the unstable decomposed translation (as they are too close) of found Fundamental matrix since we use loop-closures for the experiment. See Table.I and II for results under different datasets and Fig....

Under the well-lit condition, we observe in the Table.I that the proposed BIG descriptor generates higher errors compared to the other pixel-level features. Two main reasons cause the lower metric: i) as the BIG descriptor utilizes the barycenter of the vertex (*i.e.* superpixel regions) as the 2D coordinate, the deformation (sometimes even disappearance) of the superpixel may heavily affect the *repeatability* on images, therefore yields higher reprojection errors. The relative lower rotational errors also suggests that though the coordinates may not be accurate, but a certain level of geometric consensus does present between image pairs and can generate correct geometric information. ii) Different to traditional features which usually extract hundreds even thousands of keypoints from image pairs. The proposed method only gives descriptors of the same quantity to the superpixel regions (50-100 for each image). Insufficient quantity inherently incites erroneous estimation during the RANSAC process.

Under the day-night shift condition, the proposed BIG method vastly outperforms all local features and achieves similar rotational error against normal conditions thanks to its high robustness and the exploitation of FGSN semantic information.

NrC dataset	Reproj Error (Pixel)		Rot Error (Deg)	
	Avg	Med	Avg	Med
ORB	18.47	**7.05**	9.01	4.76
SIFT	**18.31**	14.81	6.02	**4.09**
SURF	27.93	18.50	**5.96**	4.22
BIG	50.77	41.61	7.23	5.86

TABLE I: The average and median reprojection and rotational errors of features on Newer College Dataset

RobotCar dataset (Ref vs. Night)	Reproj Error (Pixel)		Rot Error (Deg)	
	Avg	Med	Avg	Med
ORB	228.88	216.33	79.24	89.76
SIFT	621.66	213.25	71.14	79.98
SURF	514.69	341.65	67.20	69.97
BIG	**115.67**	**88.19**	**17.84**	**5.41**

TABLE II: The average and median reprojection and rotational errors of features on RobotCar Dataset between the reference condition and night condition.

II. Speed Performance

In this section, we discuss the speed performance, it comprises two parts: Generation time and matching time. Generation time is define as the time cost to generate a BIG descriptor from an image or multiple layers of information. The matching time concentrates on the searching time of a already generated descriptor in a loop-closure dataset, this factor is under the influence of scalability of the searching and matching system.

A. Generation time breakdown

Generating a BIG descriptor takes multiple parts, includes all the necessary steps to build a BIG descriptor (image acquisition, semantic generation and deep feature generation are not included as they are external of the proposed system and depends on different types of devices and implementation methods):

- **Generation of the map data**: We can further categorize this section into several main steps:
  - *Generation of superpixel*: generate superpixel from RGB image.

Author are with Univ Rennes, Inria, CNRS, Irisa, France.
Email:{xi.wang}@inria.fr
{marc.christie, eric.marchand}@irisa.fr

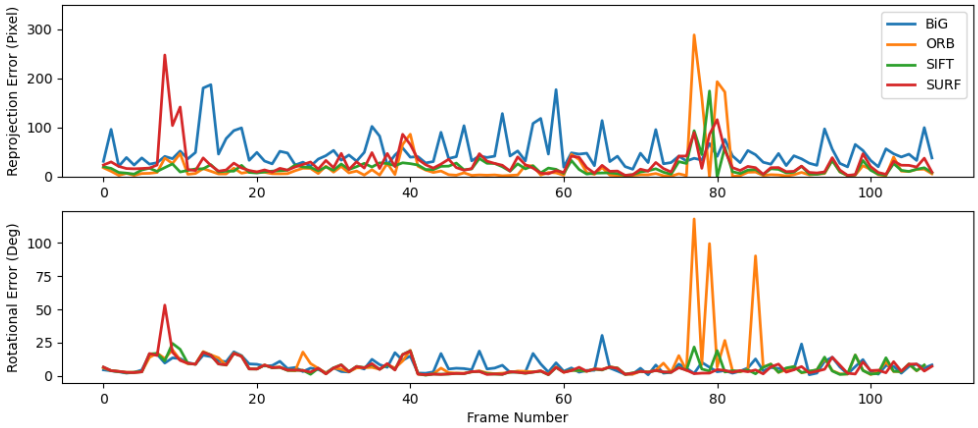

Fig. 1: The Figure of reprojection errors (top) and rotational (bottom) errors of extrated correspondences in newer college dataset, with the x-axis representing image numbers of loop-closed images, y-axis are two errors respectively. The reprojection errors of the proposed BIG descriptor (blue) are higher than other pixel-level extractor/descriptor due to its superpixel-level accuracy and insufficient descriptor quantity during the estimation procedure. However, the rotational errors compete with other methods and demonstrates the geometrical accuracy of the proposed descriptor.

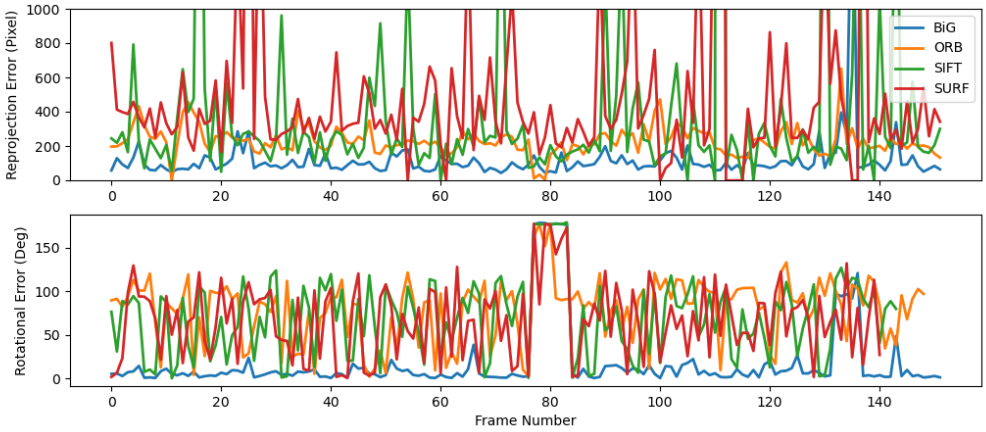

Fig. 2: The Figure of reprojection errors (top) and rotational (bottom) of extracted correspondences in RobotCar Seasonal dataset between the reference condition and night condition, with the x-axis representing image numbers of loop-closed images, y-axis are two errors respectively. Under the condition of day-night shift, the proposed BIG descriptor (blue) largely outperforms all traditional local extractors/descriptors in both metrics.

  - *Generation of edge*: building graph structure from the superpixel regions connections.
  - *Generation of the vertex*: the histogram generation is involved in this step.
- **Generation of the descriptor**: Graph embedding and binarization.

See Table. III for details of different layers information. Three observations can be made from the table: i) the time cost rises when accumulating multiple layers, but not proportionally as some fixed cost steps are only executed once for multiple layered versions; ii) the NetVLAD involved method shows lower time cost as the histogram generation step is omitted during the vertex generation stage (the outputs of NetVLAD are vectorized already); iii) the descriptor generation is very efficient against the increasing layer numbers, demonstrates the speed performance of the graph embedding and binarization process. In additional, we don't report the matching cost of *descriptors* here since the matching of Hamming descriptors are generally extremely fast in all platforms and can even reach sub-millisecond level. And about the superpixel generation, we re-use the implementation of SLIC-cpp in the experiment, but more efficient version of SLIC (such as GPU supported gSLICr [8] (reportedly 250fps) or avx2 supported fast-SLIC [1] (5-10 ms in python for ordinary size images)) or other types of

[1]Github project site: https://github.com/Algy/fast-slic

	Methods	Descriptor Generation (ms)	Graph Generation (ms)		
			SP Generation	Edge Generation	Vertex Generation
1 Layer	FGSN	36.41 ± 3.51	279.72 ± 20.38		
			132.17 ± 21.25	51.16 ± 2.70	96.39 ± 7.06
	RGB Image	35.94 ± 3.92	278.62 ± 18.64		
			135.45 ± 21.44	51.37 ± 3.24	91.80 ± 6.04
	NetVLAD	34.46 ± 3.67	173.90 ± 14.98		
			124.23 ± 15.34	49.31 ± 3.19	0.36 ± 0.08
2 Layers	Netvlad + FGSN	38.19 ± 4.26	273.56 ± 22.91		
			127.97 ± 22.65	50.46 ± 3.34	95.13 ± 8.43
	Image + Netvlad	37.14 ± 3.81	266.67 ± 19.15		
			127.70 ± 18.68	49.81 ± 2.85	89.15 ± 5.89
	Image + FGSN	41.23 ± 5.85	385.07 ± 32.96		
			140.16 ± 41.48	52.24 ± 4.41	192.67 ± 16.66
3 Layers	Image + FGSN + NetVLAD	42.52 ± 4.42	361.58 ± 23.28		
			127.89 ± 24.73	50.11 ± 2.48	183.58 ± 11.30

TABLE III: The breakdown of time cost in descriptor generation, cost is presented under the form of the mean and the standard deviation in millisecond.

superpixel generation method can be seamlessly transplanted to the BIG descriptor to further accelerate the system.

B. Matching time comparison

On the other hand, as we claimed in the paper, the main matching time cost is due to the increasing image information quantity in the database and inefficient searching scheme (*e.g.,* exhaustive linear search). In this subsection, we compare the iBoW method with a linear search scheme. Thanks to the inverted indexing technique in BoW systems, the proposed method (blue) manage to control the increasing cost against accumulating database scale, whereas the linear search method (orange) explodes rapidly.

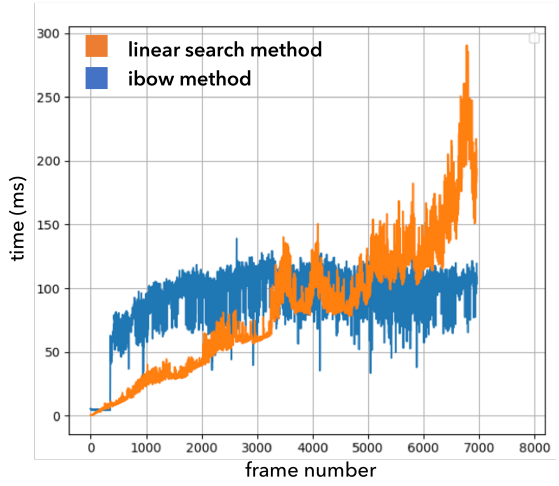

Fig. 3: Comparison between the iBoW method and linear search method used by common image retrievals. With inverted index technique, the iBoW methods (blue) shows constrained query time against the increasing scale, whereas the query time of linear search method grows proportionally (orange).

REFERENCES

[1] R. Mur-Artal, J. Montiel, and J. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Trans. on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.

[2] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *European Conf. on Computer Vision*, pp. 404–417, 2006.

[4] M. Ramezani, Y. Wang, M. Camurri, D. Wisth, M. Mattamala, and M. Fallon, "The newer college dataset: Handheld lidar, inertial and vision with ground truth," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2020.

[5] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 1, pp. 3–15, 2017. [Online]. Available: http://dx.doi.org/10.1177/0278364916679498

[6] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[7] M. Larsson, E. Stenborg, C. Toft, L. Hammarstrand, T. Sattler, and F. Kahl, "Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization," in *IEEE/CVF Int. Conf. on Computer Vision*, 2019, pp. 31–41.

[8] C. Y. Ren, V. A. Prisacariu, and I. D. Reid, "gSLICr: SLIC superpixels at over 250Hz," *ArXiv e-prints*, Sep. 2015.