2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Improving Keypoint Matching Using a Landmark-Based Image Representation

Xinghong Huang, Zhuang Dai, Weinan Chen, Li He and Hong Zhang

Abstract— **Motivated by the need to improve the performance of visual loop closure verification via multi-view geometry (MVG) under significant illumination and viewpoint changes, we propose a keypoint matching method that uses landmarks as an intermediate image representation in order to leverage the power of deep learning. In environments with various changes, the traditional verification method via MVG may encounter difficulty because of their inability to generate a sufficient number of correctly matched keypoints. Our method exploits the excellent invariance properties of convolutional neural network (ConvNet) features, which have shown outstanding performance for matching landmarks between images. By generating and matching landmarks first in the images and then matching the keypoints within the matched landmark pairs, we can significantly improve the quality of matched keypoints in terms of precision and recall measures. The proposed method is validated on challenging datasets that involve significant illumination and viewpoint changes, to establish its superior performance to the standard keypoint matching method.**

I. INTRODUCTION

Effective keypoint matching between images is a basic step for computer applications such as robotic navigation and object recognition [1]. It requires the feature detector and descriptor to have invariance properties under different types of viewing condition changes, such as viewpoint and illumination. In mobile robot visual navigation, loop closure when a mobile robot revisits a location already in the map provides critical information in constructing a map, and each loop closure hypothesis, typically generated by a highly efficient or scalable algorithm, must go through a rigorous verification step to ensure that it is not a false positive. Traditionally, this verification has depended on keypoint matching via multi-view geometry and RANSAC. Even though there exist a variety of keypoint detectors and descriptors such as SIFT [2] and SURF [3], keypoint matching with these descriptors only can experience difficulty in dealing with illumination and viewpoint changes. For visual loop closure verification using multi-view geometry (MVG), it remains a challenging issue to find a sufficient number of true matching keypoints between loop closing images.

To produce loop closure hypotheses, recently, [4] and [5] achieved state-of-the-art detection accuracy via ConvNet feature under significant environmental and viewpoint changes, showing the outstanding invariance properties of ConvNet features for matching detected landmarks between images. However, a loop closure hypothesis must go through a strict verification process, and loop closure verification has not been an active area of research, in spite of its critical role in building a practical SLAM (simultaneous localization and mapping) system.

In this paper, we present a novel method for keypoint matching that directly benefits loop closure verification. Our method first uses an object proposal method to detect landmarks in the two images of a loop closure hypothesis whose keypoints are to be matched. Then a ConvNet is used to generate the descriptors of the landmarks. We then match the landmarks before matching the keypoints in every matched landmark pair. We observe and confirm through experiments that using landmarks as an intermediate image representation can improve the performance of keypoint matching significantly because matched landmarks can effectively improve the quality of putative keypoint matches. Through experiments on challenging visual SLAM datasets involving a combination of environmental changes, we show that our method can generate a higher number of matching keypoints than the standard method that matches keypoints directly within the two images. Most importantly, our keypoint matching method results in a superior loop closure verification process according to common performance metrics.

II. RELATED WORK

Even though robot localization and mapping has been widely investigated in recent years including the issue of scalable loop closure detection [1][4][5], the issue of loop closure verification is not as widely studied. *The standard method* of loop closure verification relies on MVG to match local keypoints between the two images being considered as a loop closure. In addition to SIFT and SURF, there are various keypoint algorithms such as BRIEF [6] and ORB [7], which are designed to reduce memory and fast measurement of distance, so as to improve the performance of the feature detection and description. Furthermore, the advanced hand-crafted SIFT variants such as DSP-SIFT [8] and Root-SIFT [9] have been demonstrated to have good performance in keypoint matching [10]. Using these keypoint detection and description methods, corresponding keypoints between two images are first established by comparing their descriptor vectors and then pruned with, for example, the descriptor distance ratio test [2], where the nearest neighbor is accepted as true match when the ratio of distances to the nearest and second nearest neighbor is lower than a threshold.

All authors are with the Biomimetic and Intelligent Robotics Lab (BIRL), Guangdong University of Technology, Guangzhou, China, 510006. H. Zhang is the corresponding author (hzhang@ualberta.ca) and is also with the Department of Computing Science, University of Alberta, AB, Canada.

The work in this paper is supported by the Natural Science Foundation of China (Grant No. 61673125, 61703115), and the Frontier and Key Technology Innovation Funds of Guangdong Province (Grant No. 2016B090910003).

978-1-5386-6026-3/19/$31.00 ©2019 IEEE

Pruning can also be performed with the mutual consistency test. When there is only minor or moderate illumination and viewpoint changes, most of the true matches can be found and false matches eliminated through simple pruning techniques such as mutual consistency test. The putative matches after pruning are typically subject to MVG, in which RANSAC is used to determine the final inlier matches. Loop closure is considered true if a sufficient number of inlier matches exist between the two images of a loop closing hypothesis.

To overcome the limitation of the standard keypoint matching methods, various attempts have been made. [12] proposed a method to choose an appropriate color space in order to improve the stability of the keypoint descriptor. The specific color descriptors including the color SIFT showed a better performance than the original SIFT descriptor under illumination changes. [13] studied the performance of keypoints at different scales, and showed that those extracted at coarse scales usually provide better matching performance than those at fine scales. [14] exploited the fact that the displacement of correctly matched keypoints between two images involved in a loop closure follows a Laplacian distribution due to the special characteristic of the camera motion. One could then formulate a constraint on the putative matches to prune the outliers so that RANSAC could work properly.

In all the methods described above, putative keypoint matches are generated by matching all keypoints in one image with all those in the other. When the discrimination power of keypoint descriptors is weakened by viewpoint and illumination changes, false putative matches can be such a serious problem that they prevent the subsequent inlier detection via MVG and RANSAC from working effectively.

Exploiting the invariance properties of ConvNet landmarks that have been successfully investigated recently [5][15], we adopt a coarse-to-fine approach in our method to matching keypoints in which ConvNet landmarks in the two images are matched first before the keypoints within the landmarks are matched. Landmark-specific matching of keypoints focuses on image regions where true positives must reside - assuming landmarks are matched correctly. Landmark-specific matching of keypoints can also be expected to be less vulnerable to viewpoint and illumination changes than matching between whole images as only a small set of keypoints within the landmarks are involved in the nearest-neighbor search. We establish the superior performance of our keypoint matching method with thorough experiments.

The rest of this paper is organized as follows. We will describe our proposed method for keypoint matching in Section III. The propsed method will be evaluated experimentaly in its application to loop closure verification in Section IV. Conclusions will be drawn and future work outlined in Section V.

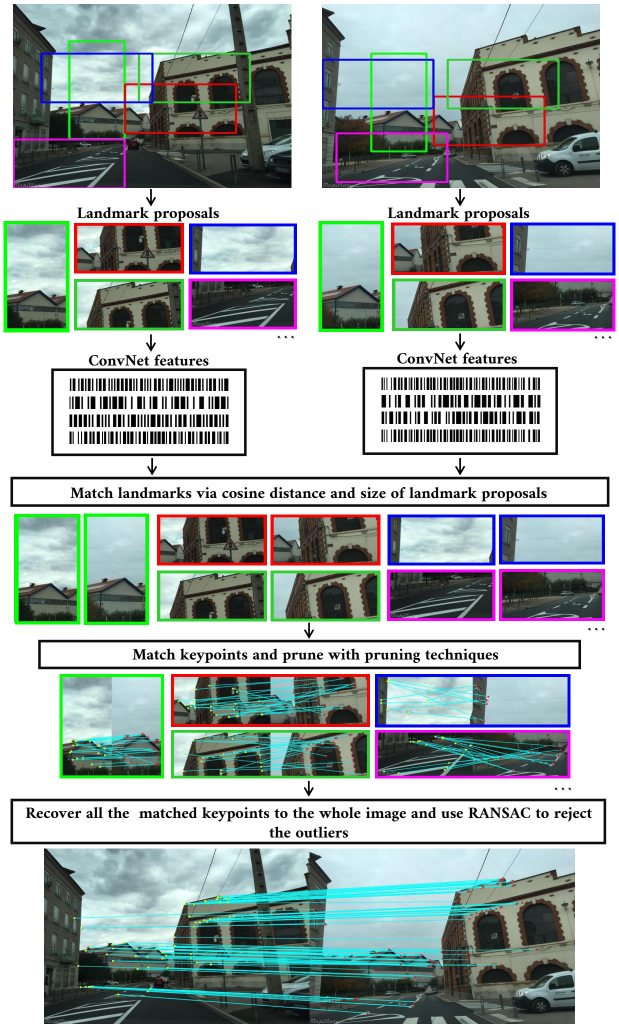

Landmark proposals
Landmark proposals
...
...
ConvNet features
ConvNet features
Match landmarks via cosine distance and size of landmark proposals
...
Match keypoints and prune with pruning techniques
...
Recover all the matched keypoints to the whole image and use RANSAC to reject the outliers

Fig. 1: Summary of the proposed keypoint matching algorithm based on matched landmarks (images from the Mapillary dataset [23]).

III. Proposed Method

In this section, we describe the four key components of our proposed method for keypoint matching based on the landmark pairs between two images, to improve the performance of visual loop closure verification. Figure 1 provides a flowchart of our proposed method, which consists of the following four key steps:

1) Extracting landmark proposals from two images whose keypoints are being matched.
2) Computing ConvNet features for detected landmarks and matching the landmarks between the two images.
3) Matching keypoints within each pair of matched landmarks and pruning the matching keypoints with the pruning technique (in order to produce high-quality putative matches).
4) Combining all the matched keypoints of the landmarks as putative matches between the two images and identifying inlier matches with MVG and RANSAC.

A. Object Proposals as Lankmarks

In this step, we use an object proposal method to generate objects and use them as landmarks. Although many object proposal methods exist and any one of them can work in our method, we use BING [16] to extract landmarks in our study because BING has been regarded as one of the state-of-the-art object detection algorithms in computer vision. In addition to BING, Edge Boxes [17] and Selective Search [18] also provide outstanding performance in object detection. However, BING is still slightly better than others according to [19]. More importantly, BING has a faster processing time than other object proposal methods, with an execution time of 24 ms per image on a laptop computer used in our study. In our experiments, as in previous studies [5][22], we extract 100 landmark proposals per image.

B. ConvNet Feature as Landmark Descriptor

Using a pre-trained convolution neural network such as AlexNet [20], we can generate the feature vector or descriptor for each landmark. AlexNet consists of five convolution layers, each of which is followed by a non-linear activation function. AlexNet also has three fully connected layers and finally a soft-max layer. Researchers have studied the various layers of the ConvNets with respect to their performance in feature description. [21] showed that the features from the mid-level layers such as the 3rd convolution layer of AlexNet (which we call Conv3 hereafter) exhibit high invariance to appearance changes. Therefore, without loss of generality, we use Conv3 for landmark description in this study. For each extracted landmark proposal, we resize it to $224 \times 224 \times 3$, the expected input size to AlexNet and from Conv3, to obtain feature vectors at 64896 dimensions.

In order to determine the similarity between landmarks l_i^a and l_j^b, which come respectively from image I^a and I^b, we calculate the cosine distance d_{ij} of their descriptors

$$d_{ij} = \frac{< l_i^a, l_j^b >}{|l_i^a||l_j^b|}$$

as their similarity. Here we use the constraint of mutual consistency, enforcing that a landmark pair as a true match only when two landmarks are the nearest neighbors of each other.

Furthermore, we also use the shape similarity of landmark proposals to filter landmark pairs. Specifically, two corresponding bounding boxes of a matched landmark pair must be sufficiently similar geometrically in the sense of:

$$max(w_i^a, w_j^b) \leqslant r \times min(w_i^a, w_j^b)$$

and

$$max(h_i^a, h_j^b) \leqslant r \times min(h_i^a, h_j^b)$$

where (w_i^a, h_j^a) and (w_i^b, h_j^b) are the widths and heights of the matched landmarks and *r* is a constant greater than 1. In our experiment, we follow the result of [22], and set *r* = 1.3.

C. Matching Keypoints Using Matched Landmark Pairs

Once ConvNet landmarks have been matched, we can proceed to generate keypoint matches within the landmarks first before pooling all the matches in the final step of keypoint matching with MVG and RANSAC. As mentioned previously, we find that the larger the bounding box of a landmark, the more false positives there will be among the matched keypoints based on just pruning technique. Therefore, we use a constraint on the size of the bounding box to limit landmark pairs to be matched as follows:

$$w_i^a \leqslant s \times w^a$$

and

$$h_i^a \leqslant s \times h^a$$

where *s* is a constant value less than 1. In our experiment, we set *s* = 0.6, and w^a and h^a are the width and the height of image I^a.

With respect to keypoint detection and description, a variety of choices exist and we choose ORB, SIFT, RootSIFT in our study, without loss of generality. ORB has several advantages over other alternatives. It has the compact representation of a binary descriptor, and is also highly efficient computationally. SIFT is a classical hand-crafted descriptor and has been widely used for keypoint matching. With respect to RootSIFT, we include it in the study because it is a recent development in hand-crafted descriptor research, and has shown excellent performance.

For each pair of matched landmarks, we treat them as image patches defined by their bounding boxes. We subsequently match their keypoints to generate a set of putative keypoint matches for the landmark pair. Note that in our experiment, we use the mutual consistency test to prune initial keypoint matches within a landmark pair.

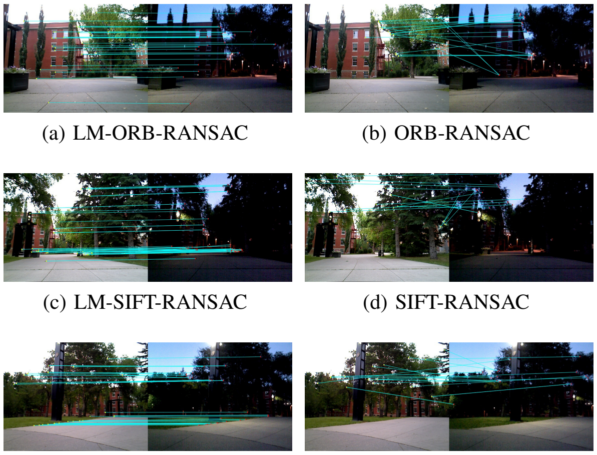
Finally, all the putative keypoint matches from the landmark-specific matching are considered as putative keypoint matches between the two images, and are subject to MVG using RANSAC to find the image-wise inliers. Intuitively, our method produces a larger number of higher-quality putative keypoint matches than the standard method, by exploiting the fact that true positive keypoint matches necessarily come from matching landmarks, and that non-landmark regions of the images tend to be poor in texture and are less likely to produce distinctive keypoints easy to be matched.

IV. EXPERIMENTAL EVALUATION

To demonstrate the performance of the proposed method, we performed experimental assessments on two datasets. In this section, the experimental procedure is firstly described in terms of testing datasets, conditions under which to compare our method and the standard method, and evaluation metrics. Then we show results with respect to the loop closure verification accuracy that reflects the effectiveness of our method.

A. Loop Closure Datasets

In our study, two popular real-world visual SLAM datasets are used, namely UACampus dataset [14] and Mapillary dataset [23]. In order to retain generality, we use a CNN-based [4] and a GIST-based [24] loop closure detection algorithm to generate potential loop closures. Note that GIST



(a) LM-ORB-RANSAC

(b) ORB-RANSAC

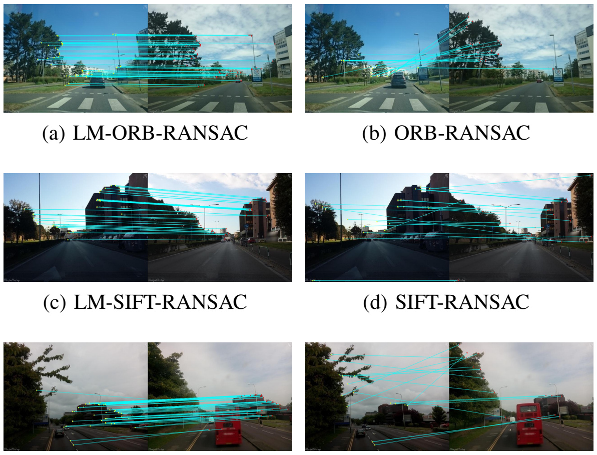
(c) LM-SIFT-RANSAC

(d) SIFT-RANSAC

(e) LM-RootSIFT-RANSAC

(f) RootSIFT-RANSAC

Fig. 2: Three positive examples case of keypoint matching test on the UACampus dataset. Under large illumination and minor viewpoint changes, the proposed algorithm finds sufficient and correct matched keypoint (left), while the baseline fails to find enough correct matches (right).



(a) LM-ORB-RANSAC

(b) ORB-RANSAC

(c) LM-SIFT-RANSAC

(d) SIFT-RANSAC

(e) LM-RootSIFT-RANSAC

(f) RootSIFT-RANSAC

Fig. 3: Another three positive examples case of keypoints matching test on the Mapillary dataset. Under significant viewpoint and moderate illumination changes, our method finds sufficient and correct putative matches (left) compared to the baseline (right).

is a popular hand-crafted descriptor used for visual loop closure detection. Then we carry out our comprehensive experiments on verifying the potential loop closing image pairs produced by these two simple algorithms, before they can be accepted as true loop closures. Details of all datasets are shown in Table I.

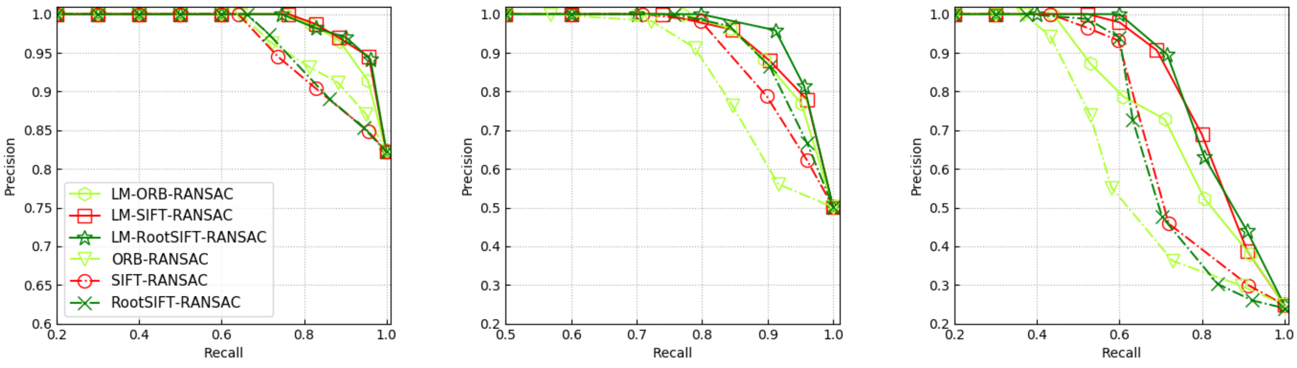
1) From the UACampus dataset, we take two subsets captured in the morning (06:20) and evening (22:15) along the same route, with maximum illumination change (see Figure 2 for an example).
2) From the Mapillary dataset, we use a subset exhibiting significant viewpoint and moderate appearance changes. In our experiments, utilizing the API interface provided by Mapillary, we downloaded 1300 image pairs with GPS information. Each image pair shows different viewpoints (see Figure 1 and Figure 3 for examples). The main purpose of this dataset is to compare methods in case of viewpoint change.
3) By UACampus-GIST and UACampus-CNN dataset, we mean the set of potential loop closing image pairs generated by the GIST-based and CNN-based loop closure detection algorithm, respectively. GIST-based algorithm produces 468 true loop closure image pairs and 102 false loop closure image pairs while CNN-based algorithm generates 630 true loop closure image pairs and only 10 false loop closure image pairs.
4) For the Mapillary-GIST and Mapillary-CNN dataset, similar to UACampus-GIST and UACampus-CNN dataset, each is a set of potential loop closure image pairs among which are 156 true loop closure image pairs and 469 false loop closure image pairs for Mapillary-GIST, and 228 true loop closure image pairs and 268 false loop closure image pairs for Mapillary-CNN.

TABLE I: Main properties of testing datasets used in our experiment. **Number** means the number of images in the corresponding subset. Note that **Potential loop closure dataset** indicates the loop closure image pairs generated by GIST-based and CNN-based loop closure detection algorithms on UACampus and Mapillary dataset.

Dataset	**Number**	**Potential loop closure dataset**		**Main changes**	
				Illmination	Viewpoint
UAcampus [14]	647 647	UACampus-GIST	UACampus-CNN	large	minor
Mapillary [23]	1300 1300	Mapillary-GIST	Mapillary-CNN	moderate	large

B. Comparison of the Proposed Method and the Standard Method

To evaluate the performance of our proposed method, we compare it with the standard method, which performs keypoint matching on the whole image directly, and then reject outliers by MVG and RANSAC. Since our experimental datasets are developed for visual robot navigation and do not have ground truth for keypoint matching, we are not able to evaluate the competing methods directly in terms of matched keypoints. Instead we compare them indirectly in terms of their loop closure verification performance for which we do have ground truth. In order to reach conclusions that are independent of keypoint detectors and descriptors, we conducted experiment on ORB, SIFT and RootSIFT, three popular and representative feature descriptors in visual SLAM community. In our experiment, we extract 500 keypoints per image. In order to make our comparative study fair, we ensure that in our method we extract approximately 500 keypoints per image across all matched landmarks. In addition the number of keypoints per landmark is the same independently of its



(a) UACampus-GIST (b) Mapillary-CNN (c) Mapillary-GIST

Fig. 4: Comparisons of competing methods in loop closure verification accuracy in terms of the precision-recall curve on four datasets of loop closure hypotheses. Note that the result of UACampus-CNN is not shown due to the almost perfect result and that the range of the y-axis of Fig. 4(a) is different from that of (b) and (c).

size, although this number does depend on the number of matched landmarks in the image. For convenience, In the rest of the paper, we use the following terms to designate the two groups of methods in the comparative study, with or without the landmark matching step are studied and evaluated:

1) ORB-RANSAC, SIFT-RANSAC and RootSIFT-RANSAC refer to the standard keypoint matching method, when it is used with ORB, SIFT, RootSIFT as detector and descriptor respectively. In all cases, a set of putative matches are first produced by the nearest neighbor search on a distance metric of keypoint descriptors and then pruned with mutual consistency test, on the whole image. MVG and RANSAC are used to reject the outliers.
2) LM-ORB-RANSAC, LM-SIFT-RANSAC and LM-RootSIFT-RANSAC refer to the proposed method paired with ORB, SIFT and RootSIFT respectively. In these cases, keypoints are matched first within each pair of matched landmarks, before the landmark-wise matched keypoints are combined as putative matches, and MVG and RANSAC are used to reject the outliers.

C. Evaluation Metric

A potential loop closure image pair is verified if the number of inliers after MVG and RANSAC is above a confidence threshold. Therefore, the performance of a keypoint matching method can be indirectly measured through verifying the loop closure hypotheses on our datasets. In our experiments, to measure performance of loop closure verification, we use the following three popular metrics:

1) Precision-recall curve is the standard criterion for assessing the loop closure verification accuracy. Define precision=TP/(TP+FP) and recall=TP/(TP+FN), where TP, FP and FN represent the number of true positives, false positives and false negatives, respectively. By changing the confidence threshold on verifying a loop closure, we can obtain a precision-recall were used curve.
2) Maximum recall at 100% precision, which indicates the performance of the evaluated method when there are no false positives, is particularly pertinent in loop closure verification.
3) Average precision (AP), a scalar indicator that reflects the overall performance of loop closure verification, is obtained by calculating the sum of all the precision values and averaging.

D. Results and Discussion

In this section, we provide the performance of loop closure verification of our method on the three keypoint features: LM-ORB-RANSAC, LM-SIFT-RANSAC and LM-RootSIFT-RANSAC, and compare them with the corresponding standard method, using the three above-mentioned evaluation metrics on four loop closure datasets. The overall results of the performance comparison are shown in Figure 4 and Table II. In general, one can easily conclude that compared to the standard method, our method shows moderate or significant advantage on all loop closure datasets and with respect to any evaluation metrics. Figure 2 and Figure 3 compare our method and the standard method qualitatively on image pairs with significant illumination and viewpoint changes. Once again, independently of descriptor used or the dataset tested, our method outperforms the standard method clearly. The details of the experimental results can be organized as follows:

1) From Figure 4 and Table II, LM-ORB-RANSAC, LM-SIFT-RANSAC and LM-RootSIFT-RANSAC show the superior performance of the proposed method when compared to the respective baseline in terms of recall value at 100% precision. For instance, in Mapillary-GIST dataset, LM-ORB-RANSAC outperforms ORB-RANSAC by 7.10%, LM-SIFT-RANSAC

TABLE II: Loop closure verification accuracy of **LM-ORB-RANSAC** vs. **ORB-RANSAC**, **LM-SIFT-RANSAC** vs. **SIFT-RANSAC** and **LM-RootSIFT-RANSAC** vs. **RootSIFT-RANSAC** in terms of maximum recall at 100% precision (Re. at 100% Pr.) and average precision (AP). The highest value with respect to each metric on each dataset is highlighted in bold. The middle values are the differences between **LM-ORB-RANSAC/ORB-RANSAC**, **LM-SIFT-RANSAC/SIFT-RANSAC** and **LM-RootSIFT-RANSAC/RootSIFT-RANSAC**.

Method	UAcampus-CNN		UAcampus-GIST		Mapillary-CNN		Mapillary-GIST	
	Re. at 100% Pr.	AP	Re. at 100% Pr.	AP	Re. at 100% Pr.	AP	Re. at 100% Pr.	AP
LM-ORB-RANSAC	**93.16%**	**99.81%**	**75.37%**	**98.10%**	**77.09%**	**96.16%**	**43.87%**	**78.64%**
	+13.83%	*+0.14%*	*+10.92%*	*+1.32%*	*+20.26%*	*+5.07%*	*+7.10%*	*+11.34%*
ORB-RANSAC	79.33%	99.67%	64.45%	96.78%	56.83%	91.09%	36.77%	67.30%
LM-SIFT-RANSAC	**89.60%**	**99.79%**	**76.23%**	**98.87%**	**74.01%**	**96.52%**	**52.26%**	**84.51%**
	+14.08%	*+0.12%*	*+12.20%*	*+2.91%*	*+3.08%*	*+1.96%*	*+9.03%*	*+10.04%*
SIFT-RANSAC	75.52%	99.67%	64.03%	95.96%	70.93%	94.56%	43.23%	74.47%
LM-RootSIFT-RANSAC	**94.44%**	**99.81%**	**74.52%**	**98.81%**	**79.74%**	**97.38%**	**60.00%**	**85.15%**
	+7.16%	*+0.10%*	*+8.35%*	*+2.68%*	*+3.97%*	*+2.29%*	*+22.59%*	*+11.57%*
RootSIFT-RANSAC	87.28%	99.71%	66.17%	96.13%	75.77%	95.09%	37.41%	73.58%

exceeds SIFT-RANSAC by 9.03% and LM-RootSIFT-RANSAC is better than RootSIFT-RANSAC by 22.59%. This observation implies that the proposed method excels in avoiding false positives. Moreover, it can be clearly observed that the proposed methods, to a certain degree, outperforms the standard image-wise keypoint matching method in terms of average loop closure verification accuracy. Although this advantage is quite moderate on the UACampus-CNN dataset due to the fact that there are only 10 false loop closure image pairs in the dataset to begin with.

2) It can be observed from Figure 4 is that the precision-recall curves produced by LM-ORB-RANSAC, LM-SIFT-RANSAC and LM-RootSIFT-RANSAC are moderately or significantly higher than those of the respective standard methods. Therefore, our method improves the loop closure verification accuracy on datasets with illumination and viewpoint changes.
3) For the qualitative results of keypoint matching in Figure 2 and Figure 3, the proposed method can produce quite sufficient inliers, in comparison to the standard image-wise keypoint matching. This is because when significant illumination and/or viewpoint changes happen, the invariance properties of the detectors and descriptors are insufficient to allow the keypoints to be matched image-wise, although with the help of matching landmarks first, successful keypoint matching is still possible. In general, landmark matching is relatively easy compared with keypoint matching due to the richer textural or semantic information available in larger image regions.

V. CONCLUSION

In this paper, we propose a simple and effective method for keypoint matching. The method uses ConvNet landmarks as an intermediate image representation to improve the quality of the putative keypoint matches, in contrast to the standard methods in which entire images are involved in a single step. The use of ConvNet landmarks generates high-quality keypoint matches by focusing on promising regions of the images and reducing the difficulty in nearest neighbor search from one whole image to the other.

Specifically, we use BING to extract landmark proposals from two images whose keypoints are being matched. Then we compute ConvNet features for detected landmarks and match the landmarks between the two images. Keypoint matching first occurs within each pair of matched landmarks and then all the landmark-specific matching keypoints are combined to serve as putative matches between the two images, subject to the final inlier detection by MVG and RANSAC.

Experimental results on multiple datasets and multiple keypoint features demonstrate the superiority of our proposed method over the standard method in visual loop closure verification. Our method has general application in keypoint matching, and is immediately beneficial to loop closure verification in visual robot navigation. Although our method does incur additional cost in detecting and matching CNN landmarks, this cost can be quite minimal since state-of-the-art loop closure detection methods already produce the landmarks that we can then use in our keypoint matching method. Our future work includes extending the experiments to other keypoint features and investigating alternative methods for producing landmarks to further improve the performance of our method.

REFERENCES

[1] Angeli, A., Filliat, D., Doncieux, S., & Meyer, J. A. (2008). Fast and incremental method for loop-closure detection using bags of visual words. IEEE Transactions on Robotics, 24(5), 1027-1037..

[2] Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), 91-110.

[3] Bay, H., Tuytelaars, T., & Gool, L. V. (2006). SURF: speeded up robust features. European Conference on Computer Vision (Vol.110, pp.404-417). Springer-Verlag.

[4] Hou, Y., Zhang, H., & Zhou, S. (2015). Convolutional neural network-based image representation for visual loop closure detection. IEEE International Conference on Information and Automation (Vol.15, pp.2238-2245). IEEE.

[5] Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., & Upcroft, B., et al. (2015). Place recognition with ConvNet landmarks: Viewpoint-robust, condition-robust, training-free. Proceedings of the 2010 Academy of Marketing Science (AMS) Annual Conference. Springer International Publishing.

[6] Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). Brisk: binary robust invariant scalable keypoints. , 58(11), 2548-2555.

[7] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2012). ORB: An efficient alternative to SIFT or SURF. IEEE International Conference on Computer Vision (Vol.58, pp.2564-2571). IEEE.

[8] Dong, J., & Soatto, S. (2014). Domain-size pooling in local descriptors: dsp-sift. 5097-5106.

[9] Arandjelović, R., & Zisserman, A. (2012). Three things everyone should know to improve object retrieval. IEEE Conference on Computer Vision and Pattern Recognition (Vol.157, pp.2911-2918). IEEE Computer Society.

[10] Schonberger, J. L., Hardmeier, H., Sattler, T., & Pollefeys, M. (2017). Comparative Evaluation of Hand-Crafted and Learned Local Features. IEEE Conference on Computer Vision and Pattern Recognition (pp.6959-6968). IEEE Computer Society.

[11] Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. ACM.

[12] Van, d. S. K. E. A., Gevers, T., & Snoek, C. G. M. (2010). Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 32(9), 1582-1596.

[13] Zhang, H. (2011). BoRF: Loop-closure detection with scale invariant visual features. IEEE International Conference on Robotics and Automation (Vol.47, pp.3125-3130). IEEE.

[14] Liu, Y., Feng, R., & Zhang, H. (2015). Keypoint matching by outlier pruning with consensus constraint. IEEE International Conference on Robotics and Automation (Vol.2015, pp.5481-5486). IEEE.

[15] Hou, Y., Zhang, H., Zhou, S., & Zou, H. (2017). Efficient convnet feature extraction with multiple roi pooling for landmark-based visual localization of autonomous vehicles. Mobile Information Systems, 2017(1), 1-14.

[16] Cheng, M. M., Zhang, Z., Lin, W. Y., & Torr, P. (2014). BING: Binarized Normed Gradients for Objectness Estimation at 300fps. Computer Vision and Pattern Recognition (pp.3286-3293). IEEE.

[17] Zitnick, C. L., & Dollr, P. (2014). Edge Boxes: Locating Object Proposals from Edges. European Conference on Computer Vision (Vol.8693, pp.391-405). Springer, Cham.

[18] Uijlings, J. R. R., Sande, K. E. A. V. D., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. International Journal of Computer Vision, 104(2), 154-171.

[19] Hosang, J., Benenson, R., Dollr, P., & Schiele, B. (2016). What makes for effective detection proposals?. IEEE Transactions on Pattern Analysis & Machine Intelligence, 38(4), 814-830.

[20] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems (Vol.60, pp.1097-1105). Curran Associates Inc.

[21] Sünderhauf, N., Shirazi, S., Dayoub, F., & Upcroft, B. (2015). On the performance of convnet features for place recognition. 4297-4304.

[22] Hou, Y., Zhang, H.,& Zhou, S. (2017). Bocnf: efficient image matching with bag of convnet features for scalable and robust visual place recognition. Autonomous Robots(9), 1-17.

[23] Mapillary, https://www.mapillary.com, accessed on May 15, 2016.

[24] Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. Kluwer Academic Publishers.