Don't Be Confused: Region Mapping Based Visual Place Recognition

Dapeng Du, Na Liu, Xiangyang Xu and Gangshan Wu

State Key Laboratory for Novel Software Technology
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing University, Nanjing 210023, China
dudp.nju@gmail.com, liunana1993@gmail.com,
xiangyang.xu@smail.nju.edu.cn, gswu@nju.edu.cn

Abstract. Visual place recognition is usually formulated as a general image retrieval problem which suffers from numerous demanding and realistic environment challenges. In this paper, we exploit the particularity of place images which can be surprisingly helpful on place recognition. Specifically, we find that images of identified places can be effectively matched by remarkable regions like building facades under limited geometry and illumination changes. Based on that observation, a novel region mapping based method is proposed to comprehensively tackle the influences caused by geometric and illumination variance as well as irrelevant interference. Given a query image, we extract remarkable regions with color constancy feature performed at processing illumination variant conditions. We leverage a two-fold transformation estimation based verification strategy dealing with geometry transformation caused by viewpoint changes for matching. The experimental results demonstrate that the proposed method is powerful for visual place recognition.

1 Introduction

Visual place recognition, which is dedicated to finding images depicting the same place via a query image of a particular street or a building [1], is of fundamental importance to many applications, such as image-based localization [2], landmark recognition [3], and loop closure detection of SLAM (simultaneous localization and mapping) in robotics [4].

Visual place recognition is often tackled with image retrieval techniques, which suffers from numerous demanding and realistic environment challenges. For example, in case (a) of Fig. 1, two images depicting totally different places improperly exhibit a good similarity due to matching of interference features. In case (b), two images of the same place appear quite different because of viewpoint and illumination changes.

In this paper, we present a novel region mapping based method to address these challenges. After an in-depth study to this problem, we find that searching remarkable regions (e.g., building facades) instead of entire images is surprisingly effective in place recognition. Further, the retrieval of remarkable regions

(a) (b)

Fig. 1. Examples of challenges in visual place recognition. (a): two images depicting different places get mismatched caused by irrelevant interference match (yellow keypoints). (b): two images depicting the same place from quite different viewpoints with illumination changes.

can be simplified as a region matching problem considering geometry and illumination changes. Towards this goal, we first employ YOLO [5] to train a building detector extracting remarkable regions from a query image. The region extraction effectively makes the recognition less influenced by common non-distinctive interference, like vehicles, billboards, trash cans and so on when measuring the similarity of places. Then we alleviate the illumination changes by illumination invariant imaging processing using color constancy feature. A two-fold transformation estimation based verification homography is performed to map query regions to reference images in database, which dedicates to dealing with viewpoint changes via geometric transformation constraint. At last, reference images in database are ranked based on quantity of matching inliers which depicts the overall similarity against query image.

To evaluate the performance of the proposed method, we evaluate our method on a public dataset. Besides, to make comprehensive comparison, we build a new challenging dataset which covers more demanding and realistic changing environments such like significant illumination variance, large viewpoint change, realistic photographing noise, and conditions with irrelevant interference. This dataset has made a very significant makeup for the existing datasets which are usually comprehensiveness scarcity on those challenges. The experimental results on both datasets show that our method outperforms competitive visual place recognition methods.

The major contributions of this paper are briefly summarized as follows.

- We propose a novel method for place recognition which exploits the particularity of place images based on remarkable regions.
- We simplify the problem as a region matching process considering viewpoint and illumination variance, which performs effectively in place recognition.
- We introduce a new dataset exhibiting challenging viewpoint and appearance variation as well as rich irrelevant interference in daily life.

2 Related Work

The aim of image retrieval is at finding as many relevant database images for a given query image as possible which also provides a meaningful context for other applications. Instead of retrieving all the relevant images, imaged-based place recognition tends to retrieve these relevant images which should be exactly the same place of the query and just one matched result is sufficient in measurement metrics. In [8], a BOF image retrieval system uses the analogy of visual words that tends to represent local features in a global feature representation manner. Many visual place recognition methods build on efficient image retrieval techniques and results rely heavily on the clustering precision of visual words [1,10].

CNN features have been proved powerful for numerous computer vision tasks, such as object classification and detection [12]. In [13], Sünderhauf et al. extracted image descriptors from the stacked output of a single CNN layer and evaluated different layers finding that the lower convolutional layers to be the relatively robust against image appearance change while higher layers to viewpoint changes. In [7], object proposal technique is leveraged to obtain patches for representing landmarks from query image and images in database and pretrained CNN features are used to calculate region similarities. However, the "landmarks" patches they get are generated from specific objectness technique, many of which inevitably contain kinds of trivial and interference objects which would affect the measurement performance. Instead, we propose to target more representative regions and simplify the problem as region matching procedure with specific features processed in stead of relying on blackbox feature representation.

3 Approach

In this section we describe how we tackle this problem. We give an overview of our framework in Fig. 2. Given an input image as query, we utilize a specific detector for extracting remarkable regions. For those extracted regions, we handle them with color constancy for the robustness of illumination invariance. Then, a twofold transformation estimation based verification strategy is performed when mapping query regions to the reference images in database. At last, we measure the overall similarity between image pairs through region matching results.

3.1 Representative region extraction

Good region extraction is very important to this problem because it eliminates the interference from background and represents the peculiarity of the place image. We leverage a powerful object detection method [5] to train a building detector. Training data is mainly from Flickr[1] in which we could obtain all

[1] https://www.flickr.com/

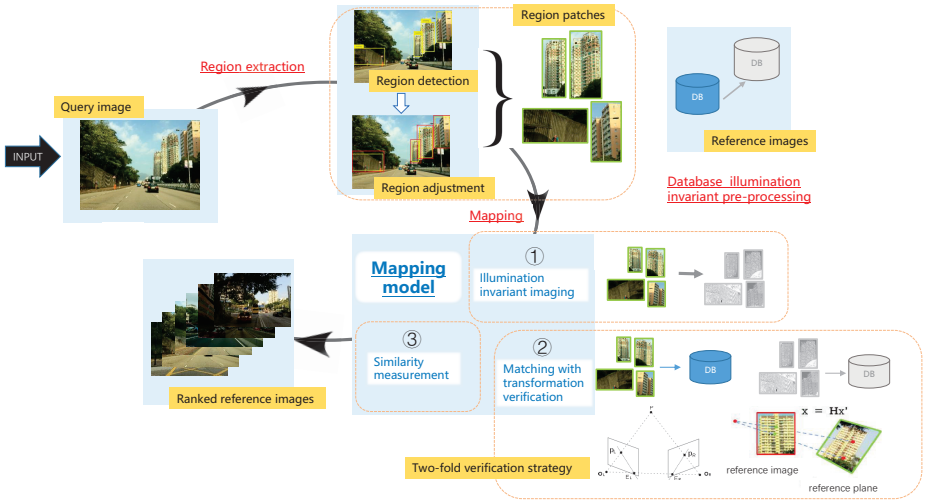

Fig. 2. The framework of our method

kinds of building facade cases. We get remarkable regions in the form of detected bounding boxes by this trained detector. According to our further observation, surroundings or peripherals of buildings play important roles in helping recognize a place. However, these bounding boxes generated for building regions often fail to envelop the whole detected objects, so we do not directly use these ones as query regions. The adjustment of bounding boxes is adapted from [6] in which we make it so that we could keep more tolerance for extension than tightness in our method (See Fig. 2, bounding box adjustment from initial ones (yellow) to adjusted ones (red)). In other words, high recall of pixels around the initial regions is more welcomed here and we also find this pixel straddling based strategy could help alleviate the error accumulation for later procedure. Remarkable regions extracted for the place are obtained then, see Fig. 2.

3.2 Illumination invariant imaging

Illumination changes make many powerful feature descriptors fail to correctly match the same place, as shown in Fig.1. It is necessary to eliminate the most of its influences for better discrimination between places. As color constancy demonstrates comprehensive relationship of an object's material property, illuminant intensity and lighting spectrum, we utilize it as a pre-processing [11] before region matching procedure. In this way, it suffers less from illumination variance and thus provides a robust condition for the region similarity measurement. Following [11], we use a one-dimensional color space $\mathcal{I}$ consisting of three sensor responses R_1, R_2, R_3 corresponding to peak sensitivities at ordered wavelengths $\lambda_1 < \lambda_2 < \lambda_3$:

$$\mathcal{I} = \log(R_2) - \alpha \log(R_1) - (1 - \alpha) \log(R_3) \tag{1}$$

The intensity of pixel x in $\mathcal{I}$ could be uniquely identified if the parameter α satisfies the following constraint:

$$\frac{1}{\lambda_x} = \frac{\alpha}{\lambda_1} + \frac{(1 - \lambda)}{\lambda_3} \tag{2}$$

Fig. 3. Using illumination invariant color space to eliminate illumination changes at different times of day. RGB images are converted to an illumination invariant color space.

The values of λ_i depend on the camera. In [11], several groups of reference values are given. Considering data distribution we use an approximate reference value here. In practice, approximation almost does not hurt matching performance. An example of producing illumination invariant color space is illustrated in Fig.3. Despite large changes in sun angle, shadow pattern and illumination spectrum between images captured at different times of day, both illumination invariant images exhibit minimal variation. We conduct this processing on all images in database with raw images preserved. Similar processing goes on extracted regions as well.

3.3 Region mapping

Viewpoint changes often lead to dramatic geometric variance. In this problem we simplified it as a plane to plane mapping procedure because buildings are generally man-made flat objects which could be approximately regarded as planar surfaces. Theoretically, no solution is available in cases that homography is not strictly applicable. However, in real applications, there is no perfect homography relation, even for planar scenes. Hence, the problem is actually derived to a minimization problem and an approximate solution is returned. Even for those buildings with curved surfaces, we find it sufficient to utilize a plane approximation since the structural variance of the surface can be ignored due to the view distance.

Specifically, we adopt a two-fold spacial estimation strategy considering the trade-off. Firstly we estimate affine epipolar geometry (i.e. the geometric relation using fundamental matrix) and the matches are listed ordered by their number of inliers. Unlike the planar homography which provides a point to point mapping, outliers might be considered as inliers since the constraint of fundamental matrix is that the correspondence must lie on the epipolar line, which is not a very strong restriction. To filter out false positives while trying to keep as many true positives as possible, a simple but effective heuristic is used in our experiment: a loose planar homography is fitted by RANSAC [15] to the inliers of the fundamental matrix and if less than 60% of these inliers are consistent with the homography then the image match is rejected. We use Hessian Affine detector [18] and SIFT descriptor [16] to extract local invariant features. For each region we make two mappings. One is from raw regions to raw reference images while the other is in the similar vein but pre-processed with color constancy before matching. Results from these two mappings will get fused in similarity measurement step.

3.4 Similarity measurement

We measure the similarity between query and referenced place images by region mapping performance. We rank the reference images in database ordered by the fused number of matching inliers. Let f_i be the function to calculate the number of inliers for the ith region extracted from query image to match the jth reference image in database. The overall similarity between this pair images is calculated by the total number of inliers N_j, as shown in Eq.3

$$N_j = \sum_i f_i \tag{3}$$

Note that f_i calculates two kinds of inliers for each region as described in Sec.3.3, i.e., from raw region to raw reference image and region-image after color constancy processing. Average weighted strategy is performed when we fuse these two kinds of inliers for robustness. In our experiment, the coefficient is set to 0.5.

4 Experiments and Analysis

In this section, we describe the experiments and analyze the results. We adopt the common evaluation metric [3,10,1], i.e., the query place is regarded as correctly recognized if at least one relevant image (within distance = 25 meters) is contained in the top N retrieved images. This has been a common place recognition evaluation metric. The percentage of correctly recognized queries is then plotted for different values of N, the so-called recall@N.

4.1 Datasets

We conduct our experiments on the public Gardens Point dataset[2] which has been widely used to evaluate place recognition methods [7]. One subset of it was recorded keeping on the left side of the walkways, while another from the right side. The dataset thus presents viewpoint changes.

Besides, considering that existing place recognition datasets only cover simple cases like near-duplicate contents and lack of challenges such as large viewpoint change, illumination variance and diverse interference, we introduce a new dataset called "MGC Places dataset". The collection source is mainly from Mapillary [3], **G**oogle Street Views and Photos captured by ourselves and this dataset covers more **C**hallenging cases. We collect 806 pictures of about 80 places manually. We also intentionally add a few interference and noisy images to our dataset. The ground truth is derived from the GPS information of images' meta data.

4.2 Compared methods and experimental settings

We compare our method with these methods as follows:
Baseline. We set the baseline according to that in [1].
Hamming Embedding with burstiness. The 64-bit SIFT Hamming Embedding (HE) [9] is proved to outperform the state-of-the-art methods when applied with burstiness normalization in the place recognition problem [19].
Coupled Multi-Index. Coupled Multi-Index (c-MI) [20] builds an effective multi-index on Hamming Embedding coupled with Color Names descriptor and is open sourced. A color codebook of 200 size is trained on independent data.
ConvNet Landmarks. In [7](Conv-landmark), Zitnick et al. extract object proposals(50 or 100) both from query images and reference images as region landmarks using [21], pre-trained AlexNet is used to extract conv3 feature to calculate similarity with cosine Euclidean distance.
Ours-ConvNet. As ConvNet-Landmarks also works with regions, we provide a variant of our method for comprehensive analysis. We extract conv3 feature for our remarkable regions with the same similarity measurement as [7] does.

4.3 Experimental Results

On the "MGC Places dataset", our method outperforms image retrieval based methods by a significant gap, as shown in Fig. 4. Specifically, we improve the recall percentage by about 15% at the best match over c-MI and HE, and 35% over baseline. This can be explained by that the comparison methods lack of effective illumination, viewpoint variance handling. The interference objects such as vehicles also affect their performance. It is interesting that though c-MI employs color cue for better performance the promotion is quite limited compared to HE.

[2] http://tinyurl.com/gardenspointdataset
[3] http://www.mapillary.com

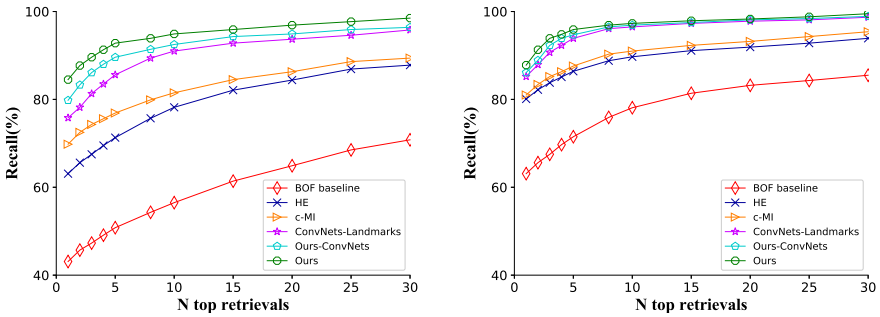

Fig. 4. Results on MGC Place dataset(left), and Garden Point dataset(right)

We think this is because local features and color cue only reduce the influences of false positive matches, and the discrimination of different appearances of the same place has limited improvement. As for ConvNet-Landmark, we conduct an extensive experiment using extracted regions from our method with pre-trained CNN features to evaluate like [7]. The extensive experiment's result shows that our method has better region extraction strategy which displays more representative and alleviates interference for recognizing a place. With specific feature processed, our method achieves a surprising performance for this task.

Comparison on Garden Point dataset presents similar results, as shown in Fig.4. The difference is that all of the comparison methods perform better than that on "MGC Places dataset". This is easy to explain since the images of the Garden Point dataset are captured sequentially, which exhibits near-duplicate scenarios in daytime. Besides, there are only 2-3 meters of camera movement, thus there are minor viewpoint changes. We note that there are some interesting failure cases for our method. These cases are kind of indoor scenarios as the walker who recorded the dataset went through an open type house during some sequences. It can be inferred that our method would work better when working on the shortlist from large image retrieval results in this condition.

Fig. 5 shows some challenging examples of place recognition showing the Top 1 results returned by our method, the ConvNet-Landmarks, the c-MI, the HE and baseline, in columns (a), (b), (c), (d) and (e) respectively. Rows from top to bottom exhibit different conditions including large viewpoint change, illumination variance, and non-distinctive interference (the last two rows, trash can and car as interference respectively). In the first example, the query has quite large viewpoint change with relevant images, thus other methods fail to handle this condition effectively. The next example shows a distinct illumination condition with intense shadow. Since features in other methods fail to do well in describing illumination invariance while we leverage color constancy, only our method and ConvNet-Landmarks get the right result. In last two examples, most methods get fake "good" matching confused by interference object in scenes, however, we overcome this challenge by matching proper regions.

5 Conclusion and Future Work

In this paper, we have presented a novel region mapping based method for image-based place recognition. By utilizing color constancy and two-fold estimation verification strategy on remarkable regions, we produce an impressive result in severe challenging conditions. Consider the shortage in existing datasets, we also introduce a new challenging dataset exhibiting extreme viewpoint and illumination variance as well as rich irrelevant interference and their combinations.

CNN based features have shown great power in this task without bells and whistles, especially the power for illumination change discrimination. We also prove that geometry transformation verification demonstrates significant in coping with viewpoint change in this problem. In the future, we will try to explore the CNN based spatial constraints in geometry transformation for better performance.

Acknowledgments. This work is supported by the National Science Foundation of China under Grant No.61321491, and Collaborative Innovation Center of Novel Software Technology and Industrialization.

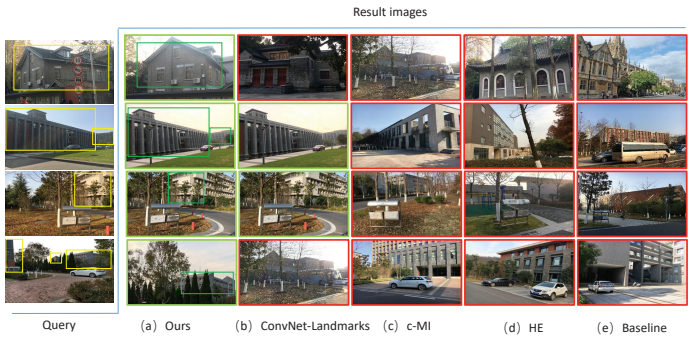

Fig. 5. Challenging examples from MGC dataset with Top 1 results displayed. Positive results are labeled with green frames while negative ones with red. Regions detected in queries are labeled with yellow frames.

References

1. Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi.: Visual place recognition with repetitive structures, in CVPR, 2013.
2. Grant Schindler, Matthew Brown, and Richard Szeliski.: City-scale location recognition, in 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2007.

3. David M Chen, Georges Baatz, Kevin Köser, Sam S Tsai, Ramakrishna Vedantham, Timo Pylvänäinen, Kimmo Roimela, Xin Chen, Jeff Bach, Marc Pollefeys, et al..: City-scale landmark identification on mobile devices, in CVPR. IEEE, 2011.
4. Mark Cummins and Paul Newman.: Fab-map: Probabilistic localization and mapping in the space of appearance, The International Journal of Robotics Research, 2008.
5. Redmon, Joseph and Divvala, Santosh and Girshick, Ross and Farhadi, Ali.: You only look once: Unified, real-time object detection, in CVPR. IEEE, 2016.
6. Chen, Xiaozhi and Ma, Huimin and Wang, Xiang and Zhao, Zhichen.: Improving object proposals with multi-thresholding straddling expansion, in CVPR. IEEE, 2015.
7. Niko Sunderhauf, Sareh Shirazi, Adam Jacobson, Feras Dayoub, Edward Pepperell, Ben Upcroft, and Michael Milford.: Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free, Proceedings of Robotics: Science and Systems XII, 2015.
8. Josef Sivic and Andrew Zisserman.: Video google: A text retrieval approach to object matching in videos, in ICCV. IEEE, 2003.
9. Herve Jegou, Matthijs Douze, and Cordelia Schmid.: Hamming embedding and weak geometric consistency for large scale image search, in ECCV. Springer, 2008.
10. Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt.: Image retrieval for image-based localization revisited., in BMVC, 2012.
11. Will Maddern, Alex Stewart, Colin McManus, Ben Upcroft, Winston Churchill, and Paul Newman.: Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles, in ICRA, 2014.
12. Sharif Razavian, Ali and Azizpour, Hossein and Sullivan, Josephine and Carlsson, Stefan.: CNN features off-the-shelf: an astounding baseline for recognition, in CVPR Workshops, 2014.
13. Sünderhauf, Niko and Shirazi, Sareh and Dayoub, Feras and Upcroft, Ben and Milford, Michael.: On the performance of convnet features for place recognition, in IROS, 2015.
14. Richard Hartley and Andrew Zisserman.: Multiple view geometry in computer vision, Cambridge university press, 2003.
15. Martin A Fischler and Robert C Bolles.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM, 1981.
16. David G Lowe.: Distinctive image features from scale-invariant keypoints, IJCV, 2004.
17. Relja Arandjelović and Andrew Zisserman.: Three things everyone should know to improve object retrieval, in CVPR. IEEE, 2012.
18. Krystian Mikolajczyk and Cordelia Schmid.: Scale & affine invariant interest point detectors, IJCV, 2004.
19. Relja Arandjelović and Andrew Zisserman.: Dislocation: Scalable descriptor distinctiveness for location recognition, in ACCV. Springer, 2014.
20. Liang Zheng, Shengjin Wang, Ziqiong Liu, and Qi Tian.: Packing and padding: Coupled multi-index for accurate image retrieval, in CVPR, 2014.
21. Zitnick, C Lawrence and Dollár, Piotr.: Edge boxes: Locating object proposals from edges, in ECCV, 2014.