IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED BIG DATA ANALYSIS FOR VEHICULAR SOCIAL NETWORKS

Received April 28, 2018, accepted May 24, 2018, date of publication June 5, 2018, date of current version June 26, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2844221

Intensity Filtering and Group Fusion for Accurate Mobile Place Recognition

MAO WANG[1], EN ZHU[1], QIANG LIU[1], (Member, IEEE), YONGKAI YE[1], YUEWEI MING[1], AND JIANPING YIN[2]

[1]Department of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China
[2]School of Computer Science and Network Security, Dongguan University of Technology, Dongguan 523000, China

Corresponding authors: En Zhu (enzhu@nudt.edu.cn) and Jianping Yin (jpyin@dgut.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1003203 and in part by the National Natural Science Foundation of China under Grants 61672528, 61773392, and 61702539.

ABSTRACT Mobile place recognition targets at matching query images captured by mobile devices with database images collected from vehicle-mounted cameras, such as Google street view panoramas, which plays an important role in many applications. However, current solutions deriving from image retrieval suffer from the problem of low precision on top results, which significantly challenges their usability. By investigating the state-of-the-art approaches, we find that the bad illumination significantly affects initial results, and these initial results are correlative in both spatial location and visual content, which can be utilized for further improvement. In this paper, we propose an effective approach to rerank initial top-ranked results to improve the recognition recall. First, initial retrieval results with low intensity are filtered as they usually depict irrelevant places with dark background. Second, the correlation between top-ranked results is modeled as a reciprocal neighborhood graph by jointly considering spatial location and visual similarity. With the graph, the initial results are reranked based on voting similarity from the query and reciprocal neighbors. In this way, the underlying structure of initial retrieval results is exploited for refining. Experimental results on the public Tokyo 24/7 and San Francisco landmark datasets demonstrate that the proposed approach can achieve persisting improvement of recognition recall over the state-of-the-art approach.

INDEX TERMS Place recognition, image-based localization, image matching, deep feature.

I. INTRODUCTION

Mobile place recognition [1], [2], i.e., given a photo captured by a mobile device and then retrieving images depicting the same place from gallery images with geo-annotations, is a critical part for many applications, such as augmented reality [3], [4], image annotation [5], image-based localization [4], [6], [7], and robotics [8], [9]. Based on the meta-information of matching result images, such as spatial position or text annotation, the query photo can be accurately recognized. In this way, we can locate the position of query images captured by uses' mobile cameras or downloaded from the social network with the aid of geo-tagged database images. As mobile devices becoming a convenient entrance to sense the physical world and interact with the Internet [10], mobile place recognition will attract more attention in increasing application occasions.

Mobile place recognition is a challenging problem as images are captured in an un-constrained environment, such as different viewpoints, large variation of illumination, and occlusion by cars or pedestrians. Current mobile place recognition relies on the advance of image retrieval. The photo to be identified is used to retrieve relevant images by querying the database in the framework of Bag of Visual Word (BoW) model [13]. Popular approaches developed for image retrieval can be employed to improve recognition accuracy, such as hamming embedding [14], [15], burstiness weighting [16], [17], spatial verification [18] and query expansion [19], [20]. Recently, deep-learning based approaches [12], [21] have also been developed for efficient retrieval and recognition.

After retrieving results, the query photo can be regarded as successfully recognized if there exists one correct image in the top N ranked list, which is the evaluation metric commonly used for place recognition, i.e., recall@N. A high recall can be achieved by current popular approaches, such as recall@50 and recall@25 are all about 90% in [14] and [12] for popular benchmark datasets respectively, which means most queries can be correctly recognized by the

top 50 or 25 ranked results. However, as pointed out in [18], high precision is also a critical performance metric for some applications, such as in mobile devices, where relevant results should be ranked as top as possible. Due to the requirements of low responsible time and the limited resources of mobile devices, it is important to figure out the performance of image retrieval based methods in terms of recall@N with small value of N, such as $N = 1$. According to experiments for the Tokyo 24/7 dataset [11] in [12] and the Pittsburgh dataset [16] in [14], the evaluated results of recall@1 from the two state-of-the-art methods are both about 70%. There still exists much room to be improved for mobile place recognition.

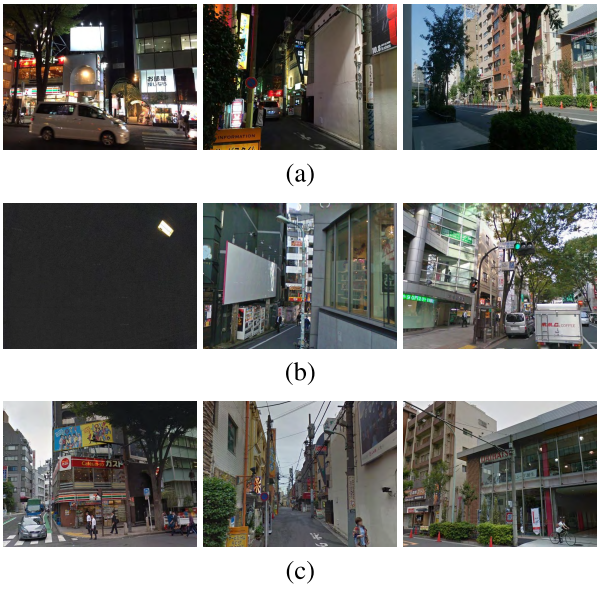

FIGURE 1. Illustration of three examples that NetVLAD [12] fail to return relevant results at top 1 on the Tokyo 24/7 dataset [11], while the proposed approach can re-rank them to top 1 position. As trained on images captured at daytime, the NetVLAD feature is confused by the dark background and similar but irrelevant buildings, leading to a bad performance for query photos with occlusion or captured at sunset or night. (a) Mobile phone Queries in Tokyo 24/7 [11]. (b) Top-1 results returned by NetVLAD [12]. (c) Top-1 results after reranking by the proposed approach.

In place recognition, the representation of images is either from conventional features, such as the BoW [13], [16] or VLAD [11], [22] descriptors deriving from hand-craft local features, or extracted by deep networks that are directly trained in an end-to-end manner, such as the NetVLAD feature [12]. In this paper, we focus on the state-of-the-art feature NetVLAD as this feature shows excellent performance in place recognition over others, such as the BoW and dense VLAD. By investigating the results returned by NetVLAD as illustrated in Fig. 1, we can find that query photos taken at night will match database images with dark background or images that have the same layout but different details. We argue that the training images of NetVLAD are captured at daytime and weakly labeled. Therefore, the learned feature will be confused by the dark background that contributes dominant similarity, or frequently appeared objects, such as trees, pedestrians and windows in street view images. However, we have observed that these images with dark background have a different intensity distribution from images captured at daytime. In addition, the initial results of current methods are not isolated from each other as they are collected by surveying vehicle cruising around streets. Thus, some of them have nearby spatial locations and depict the same scene which can be exploited for further improvement.

In this paper, an intensity filtering approach is proposed to remove irrelevant results with dark background. Then, the spatial adjacent relationship and visual similarity of the top-ranked database images are exploited to discover their underlying correlation in the framework of reciprocal neighborhood graph. After that, the most relevant result of the query is selected and reranked to the top 1 position of the final result list. In this way, we can select the positive result with a high confidence and avoid degenerating the performance of initial results returned by the NetVLAD feature. The proposed approach combines the powerful NetVLAD feature and extra information of database images to improve precision of initial results. It can be easily implemented and plugged into current methods as a post-processing step.

The contributions of this paper are summarized as follows: firstly, we propose an approach to filter irrelevant results with low intensity for queries captured under bad illumination. Then, the correlation among top-ranked images is exploited to refine initial results by considering spatial and visual information. We evaluate the proposed approach on the public Tokyo 24/7 dataset which contains query images captured in different conditions and is challenging for current approaches, and the San Francisco landmark dataset [6] which is a city-scale dataset for place recognition. Experimental results over the Tokyo 24/7 dataset demonstrate the overall recall improvement by the proposed approach, and specifically the recall@1 is improved from 71.4% to 76.5%. For the San Francisco landmark dataset, the recall@1 is improved from 72.4% to 74.7%.

The rest of this paper is organized as follows. The related work of mobile place recognition is reviewed in Section II. The proposed intensity filtering strategy is introduced in Section III. The proposed group place fusion is presented in Section IV. Experiments are demonstrated in Section V. We give conclusion remarks in Section VI.

II. RELATED WORK

Benefit from the public available datasets, such as the San Francisco landmark [6], Pittsburgh [16], Tokyo 24/7 [11], recent years have witnessed a rapid progress in mobile place recognition [12], [14], [18], [23]. The previous work can be divided into two categories: i) local feature based approaches deriving from image retrieval, and ii) global feature based approaches that use hand-craft features or descriptors trained by deep networks. Besides powerful features from image retrieval, extra information of images such as geo-tags or

3D models, is also utilized to improve the recognition performance.

A. LOCAL FEATURE BASED APPROACHES

Mobile place recognition is usually handled in the framework of image retrieval. Popular retrieval models can be employed in this field, such as Hamming Embedding (HE) [14], dense Vector of Locally Aggregated Descriptor (VLAD) [11] and Aggregated Selective Match Kernel (ASMK) [15]. These models are derived from the BoW, which involves extracting local features such as SIFT [24], quantizing local features into visual words among a pre-trained visual vocabulary and matching database features by inverted indexes. Based on these retrieval models, many approaches are proposed to improve recognition accuracy by handling different problems in place recognition, such as burstiness weighting and improving the distinctiveness of local features.

Burstiness problem [25] is a common issue in retrieving building images and place recognition, which refers to some visual elements in an image, such as windows and bricks, appear frequently. These visual elements will dominate similarity measure and result in a high similarity for irrelevant images. To tackle this problem, adaptive weighting is proposed in [16] to modify the BoW representation by considering burst features. Sattler *et al.* [18] discovered that the same problem can also appear after spatial verification [26], and a strategy of burstiness weighting based on geometric information is proposed. Local features with the same visual word will be aggregated in ASMK [15] to discount the influence of the burstiness problem. The inter- and intra- image burstiness weighting [25] developed in image retrieval are also used for improving accuracy of place recognition [14], [18]. To adapt the different distinctiveness of local features, a density estimating approach is proposed in [14] based on the HE retrieval model.

B. GLOBAL FEATURE BASED APPROACHES

Mobile place recognition needs to match street view images containing rich details, which is slightly different from matching local regions of landmarks or objects in image retrieval. The global feature can also be an effective overall description of images, such as GIST [27] originally used in scene recognition and VLAD deriving from local features. In [23], GIST feature is utilized as a complementary clue for the local feature based retrieval to fuse global and local information. Dense VLAD is used in [11] to handle the large changes of appearance between query and database images. Departing from the traditional hand-craft feature to represent images, recent approaches rely on the popular deep learning to learn a fine-tuned feature representation based on the large amount data available on the Internet. In [12], street data collected from Google Time Machine is used to train a convolutional neural network of NetVLAD feature.

C. EXPLOITING EXTRA INFORMATION

Contrast to crawling database images from the Internet in image retrieval, database images in place recognition are usually collected by surveying vehicles or downloaded from Google Street View [6], [11], [16]. Besides the geo-tags, there may exist building ID, depth or 3D model information of street view images. In addition, as collected from adjacent streets, these images are highly correlated both in their spatial positions and the content they depicting. The extra information can be a complimentary of local or global features and exploited to improve recognition performance [6], [14], [18], [23].

The location information is commonly used to restrict candidate results by only considering database images fall into the same grid of the query [6], or clustering spatial verified database images for geometric burstiness weighting [18]. To avoid returning same false results and improve diversity, database images located close are aggregated by unique landmark suggestion [14] or spatial non-maximal suppression [11], [12], [28]. Depth-maps of panorama images are utilized to synthesis different views of images for accurate matching in [11]. The structure of database is modeled as a graph by considering visual similarity between images to improve recognition performance in [28]. The relationship of reciprocal nearest neighborhood [29] among database images is used to fuse retrieval results by local and holistic features in [23].

D. CROSS DAY-NIGHT RECOGNITION

Recently, the variation of illumination has been noticed as a problem to decrease the description ability of local features and result in bad performance in place recognition and image matching [6], [30]. In image retrieval community, this problem is ignored as popular landmark datasets, such as the Oxford5k [26], Paris6k [31] and World5k [32], consist of query images that are almost captured at daytime or under proper illumination. However, in mobile place recognition, the bad illumination such as low contrast and dark background, should be carefully handled.

In [30], popular local features in place recognition are evaluated in matching day-night images and the result shows that the large illumination change will severely affect matching performance. In [6], the shadow casted from buildings to other buildings will lead to a low contrast image, which needs histogram equalization to enhance contrast of images before feature extraction. In order to improve the matching accuracy between day and night images, dense local features with interval pixels are employed in [11] to replace previous sparse local features extracted around key-points. From the qualitative examples demonstrated in project homepage of [12], cross day-night recognition is still a problem that should be considered.

III. INTENSITY FILTERING

For some query images captured at night, we have found that the dark background will dominate the similarity measure. As illustrated in Fig. 2, some top-ranked results are images collected in tunnels, at underground parks or under flyovers, which are not the desired results. Moreover, for query images captured by users at night, we would prefer to return candidate images with bright background, such as images captured at daytime with clear details, to facilitate the final judge. In this way, users can easily identify whether the returned result is correct. Based on this intuition, we propose a simple yet effective approach to filter initial irrelevant images with low intensity.

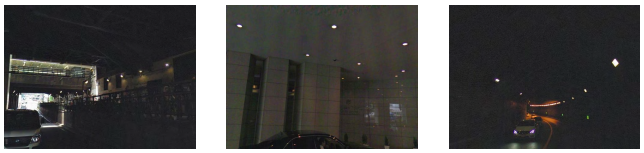

FIGURE 2. Some top-1 ranked false results with dark background returned by the NetVLAD based approach on the Tokyo 24/7 dataset.

We employ the HSI color space [33] to analysis the property of different images, where different colors are represented by three components **h**ue, **s**aturation and **i**ntensity for simulating the human visual system. The intensity component is used to describe the brightness of images, which can be utilized to distinguish dark images from bright images. As there is no existed approaches to directly distinguish dark images from bright images, we need to i) collect different types of images and ii) find the boundary of intensity for different types of images. To tackle this problem, we use query images of the Tokyo 24/7 dataset [12] that are not used for evaluation. Those images captured in different time can be utilized to verify the intensity distribution for dark and bright background images, and determine the boundary to distinguish them. There exist 1125 images in the query dataset but only 315 images are used for final evaluation, and the remains are also captured at the same place with three different time, as illustrated in Fig.3. For each image, we calculate the average intensity and analysis the variation from different groups based on the captured time. In details, the average intensity of a color image is given by

$$\text{avg\_intensity} = \frac{1}{3WH}\sum_{y=1}^{H}\sum_{x=1}^{W}\sum_{c}^{\{R,G,B\}} I(x,y,c) \qquad (1)$$

where $I(x, y, c)$ is the pixel value at position $[x, y] \in [W, H]$ of channel $c \in \{\text{Red, Green, Blue}\}$.

For images captured at daytime, sunset and night, their distributions of intensity, i.e., the mean and standard variation, are illustrated in Fig. 4. Each type of images seems to have a unique distribution of intensity, and images captured at daytime have a larger average intensity value than images captured at night. We only compare images captured at daytime and night due to the un-stable illumination

(a) (b) (c)

FIGURE 3. Queries captured at the same location but at different time. The Tokyo 24/7 dataset consists of 315 queries with 105 locations at three different time for evaluation. (a) daytime. (b) sunset. (c) night.

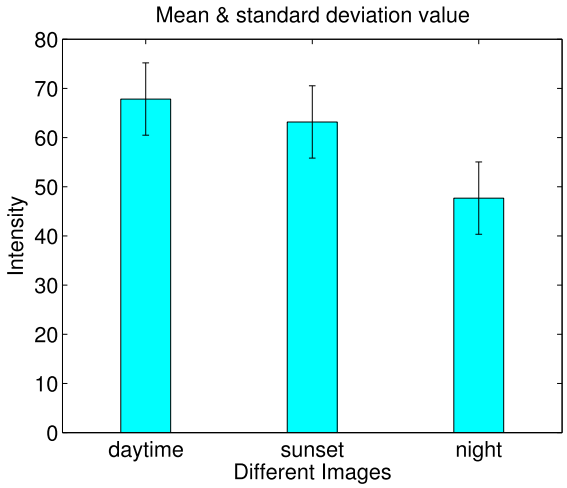

FIGURE 4. The mean and standard variation of intensity for different types of images.

condition of images captured at sunset. As implied by Fig. 4, the boundary to distinguish dark images from bright images is about 50, and we will further determine it by experiments. Even though images captured by mobile devices at night may be different from dark street view images captured by surveying car at daytime, such as the existence of light from street lamps, the proposed approach can filter irrelevant images effectively in our experiments.

IV. GROUP PLACE FUSION

Unlike general image retrieval, database images in mobile place recognition are not isolated from each other. We try to consider the underlying structure between top-ranked results returned by current state-of-the-art approaches and the reciprocal neighborhood relationship among the query and database images for further improvement.

A. SPATIAL CLUSTERING

As database images are usually captured by surveying vehicles on streets, their spatial locations are adjacent, which can be utilized to exploited the relationship between initial results. Furthermore, we have noticed that results returned by current approaches contain some relevant images depicting the same street view. As we can see from Fig. 5, these initial results returned by NetVLAD consist of some similar but irrelevant street view images and relevant images are ranked

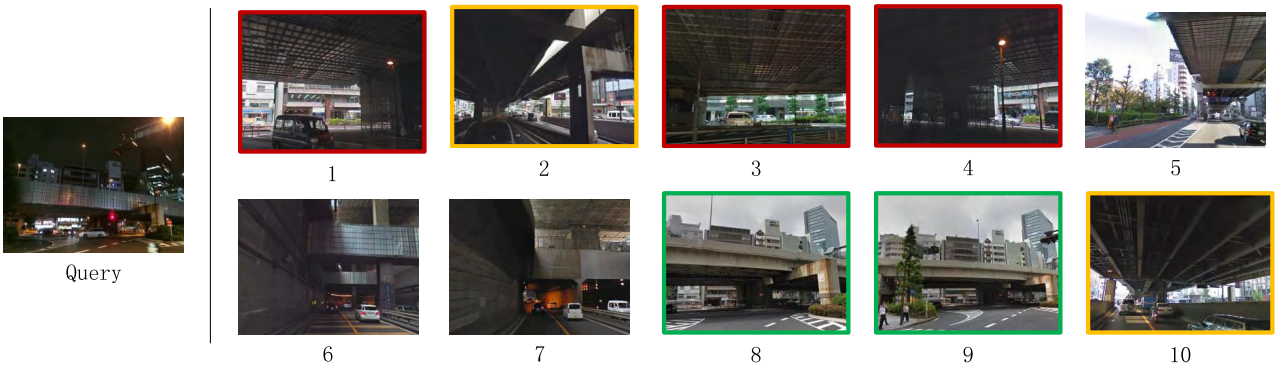

FIGURE 5. Top 10 results of a query. The illustrated results contain some spatial adjacent database images which are labeled by the rectangles with the same color, result (1,3,4), (2,10), (8,9). Due to the low contrast and unclear details, most of these results are irrelevant except result 8 and 9. These results are correlated both in spatial location and image content. We concentrate on reranking initial results by exploiting the relationship between them.

at 8 and 9. Among these results, there exist three groups depicting the similar scenes, group 1: result {1,2,4}, group 2: result {2,10} and group 3: result {8, 9}. Departing from the similar scene depicting in each group, we try to investigate the relationship of their spatial locations.

We compute the pairwise distance of top-ranked results and use the popular distance metric in place recognition to judge whether two images are spatially adjacent, i.e., if the distance of their locations is within 25 meters. Inspired by the similar idea in [16] and [34], we propose to use the connected component analysis to divide initial results into different groups. Firstly, a graph $G =< V, E >$ will be constructed, where V is the vertex set that represents top-ranked result images, and E is the edge set represents adjacent relationship among them. If result image I_i and I_j are adjacent in geo-locations, then there exists an edge e_{ij} in the graph between vertex v_i and v_j. After constructing the graph, the connected component analysis is performed on the graph to find connected groups, which contain vertexes represent images with adjacent locations. We refer to these groups as adjacent groups.

As we can see from Fig. 6, vertexes enclosed by the same ellipse consist of the same group, and others are isolated. These images depicting similar scenes are indeed spatial adjacent, which can be explained by that they are collected at nearby places. If two images are far way from each other, they cannot contain the same building or depict the same scene. The spatial adjacent of images can lead to similar content of images, which will be exploited to explore the correlation between top-ranked result images.

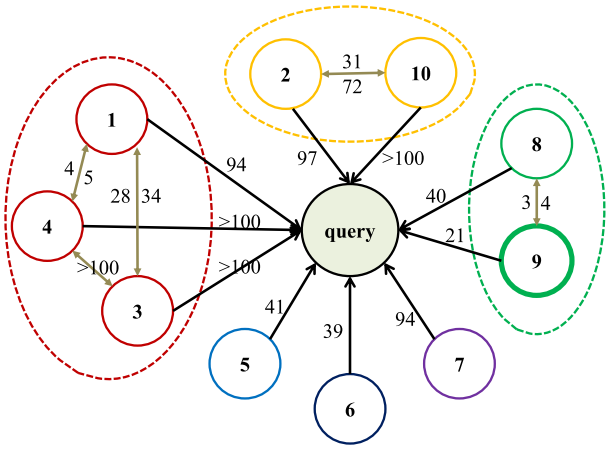

FIGURE 6. Reciprocal neighborhood graph built from the result in Fig. 5, where spatial adjacent results are enclosed by the same ellipse. The query q is in the center, and initial result $\{d_i\}$ is represent by circle around the query with number indicates its rank. The rank of the query image by using database images as query is attached with arrows. Single-direction arrows mean $rank_d(q)$, i.e., the rank of q when querying by result d, and bi-direction arrows mean $rank_{d_i}(d_j)$ and $rank_{d_j}(d_i)$, where d_i and d_j are initial result images in the same adjacent group.

B. RECIPROCAL NEIGHBORHOOD GRAPH

To discover the underlying structure of database images, the reciprocal neighborhood relationship [23], [29], [35] is employed as the similarity metric, which means two images are nearest neighbors for each other. As we can see from Fig. 6, when using the top-1 result as a query to retrieve all other images, the original query image is just ranked at 94, not in the top. The low rank implies the top-1 result may be irrelevant. This phenomenon can be explained by that the similarity between the query and result 1 are mainly contributed by the dark part which contains less discriminate information. But for result 8 and 9, the rank of the query are in the top (rank 40 and 20) by querying with result 8 and 9 respectively, which means they may be correct results of the query.

As the similarity score between images can be heavily influenced by the change of illumination in different time and seasons, or occlusion from cars and pedestrians, the result rank deriving from similarity measure on the query side may not be reliable. Similar to the work in [23] and [35], the reciprocal neighborhood relation is employed to model the similarity between query and top-ranked results. Firstly, we define the top-k nearest neighbors $N_k(q)$ of a query I_q are the top k results in the sorted result list when using I_q

to retrieve. Then, the reciprocal neighborhood relation $R(q, d)$ of I_q and I_d is defined as both I_q and I_d are the other's top-k nearest neighbors when retrieving with them respectively. If the rank of I_d in the sorted result list when retrieving by I_q is represented by $rank_q(d)$, then $N_k(q)$ and $R(q, d)$ is defined as follows.

$$N_k(q) = \{d|rank_q(d) < k\} \tag{2}$$

$$R(q, d) = d \in N_k(q) \wedge q \in N_k(d) \tag{3}$$

For the query q, its candidate results with reciprocal neighborhood relationship is represented as

$$R_k(q) = \{d|d \in N_k(q) \wedge q \in N_k(d)\} \tag{4}$$

Besides exploiting relationship between the query and initial results, the relationship between initial results among the same adjacent group is also explored as the adjacent location can lead to a high similarity measure. For example, in Fig. 6, result 1 and result 4 share a high correlation as implied by their neighborhood rank (rank 4 and rank 5), the same as result 8 and result 9. As illustrated in Fig. 5, result 1 and 4, and result 8 and 9, indeed share similar street views. However, for result 1, 3 and 4, the query is ranked very low when using them as query, which implies they are not reliable results.

Unlike the previous work in [23] that needs to compute the reciprocal neighborhood relationship of all database images offline, we restrict the reciprocal neighborhood relationship in the same adjacent group as the images depicting similar scenes should be close located. In practice, we save the top-100 similarity scores and corresponding database IDs for each database image by offline computation. When the similarity score between the query and a database image is available, the rank of the query when retrieving with the database image can be easily acquired with the previous saved information.

C. GROUP FUSION

By exploiting the spatial location information and reciprocal relationship between initial results, we can construct a graph as illustrated in Fig. 6. From the graph, we find some initial top-ranked results, such as result 1 and result 2, may not be relevant results for the query as their low reciprocal rank. However, both result 8 and result 9 have a high rank to the query, and they are highly correlated both in spatial location (in the same group) and reciprocal neighborhood relationship (high reciprocal rank for each other). Therefore, we can discover that result 9 may be a relevant result of the query. Based on the above observation, we try to fuse the reciprocal relationship and spatial adjacent information to refine initial results. As the fusion is performed on adjacent groups, we refer to it as group fusion of places.

As a low rank implies a less reliable result, we cut off some edges in the Fig. 6 by thresholding edges with rank larger than 100. The final reciprocal neighborhood graph is illustrated in Fig. 7. Based on the graph, we can exploit the

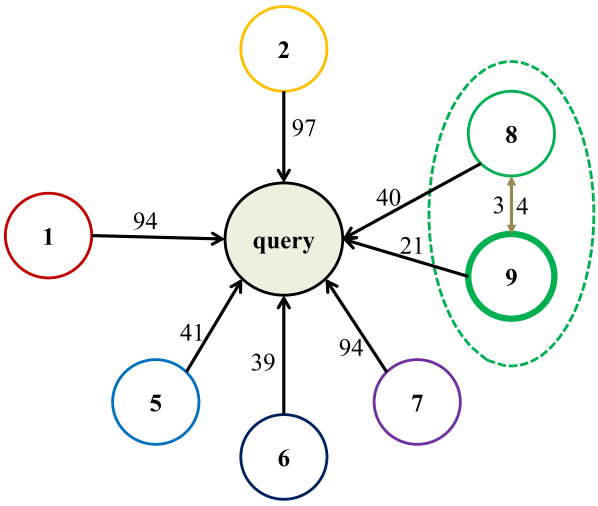

FIGURE 7. Final reciprocal neighborhood graph deriving from Fig. 6, where only candidate database images with a reciprocal rank less than 100 for the query are kept. The initial high rank group, result {1,3,4} in Fig. 6 has only result 1 left. However, for another group, result {8,9} is kept, and the two results are both located close and similar in visual content as the high reciprocal rank, indicating reliable results of the query.

relationship among initial results and rerank relevant results to the top. Similar to [35], we define similarity between two images implied by the the reciprocal rank as follow:

$$S_q(d) = \frac{1}{(rank_q(d) + rank_d(q))} \tag{5}$$

where $rank_d(q)$ is the rank of query image I_q when retrieving by the result database image I_d. If I_d is the ith result of the query I_q, then $rank_q(d) = i$. By considering voting score of reciprocal neighbors from the same adjacent group, the rerank similarity score for each database image is given by:

$$S'_q(d) = \sum_{d_i \in R_k(q) \wedge d_i \in R_k(d)} \frac{1}{(rank_q(d_i) + rank_{d_i}(q))} \tag{6}$$

For these top-ranked result images, we only consider reciprocal neighbors that are spatially adjacent to them, so the constrain $d_i \in R_k(d)$ indicates d_i and d are in the same adjacent group, i.e., the same ellipse in Fig. 7.

After acquiring the rerank score for each initial result, we select the one with max similarity score and rank it to the top 1 position of result list. If there is a tie, such as there exists a fully connected subgraph, we select the one with higher similarity score to the query, i.e., the $S_q(d)$. For the initial adjacent group result {1,3,4} in Fig. 6, there only exists one connection between result 1 and the query in Fig. 7, but there exist two connections between the query to the group result {8,9}. Based on the rerank score $S'_q(d)$, the most relevant results for the query is result {8,9}. We select result 9 and rerank it to the top as it has a higher similarity score $S_q(d)$ than result 8 ($1/(9 + 21) > 1/(8 + 40)$).

The proposed approach is inspired by the work in [23] and [35], but different from them. In [35], located objects in database will issue new queries in a way similar

to query expansion [19], [20] to filter irrelevant results and discover new relevant results. The proposed approach concentrates on reranking the initial top-ranked results by exploiting their correlation. In [23], the reciprocal graph is built from the query and expanded in multiple layers based on the similarity between all images. However, we take the spatial position into consideration and restrict to exploit relationship between initial top-ranked candidate images.

V. EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL SETUP

1) TOKYO 24/7 DATASET [11]

The dataset consists of 315 query images generated by capturing street views at different locations at daytime, sunset and night respectively. The database contains 75984 images generated by 6332 panorama images from Google Street View, 12 views for each place. This dataset is very challenging as both the query and database images undergo occlusion by cars, pedestrians and trees, view-point change and large variation of illumination. Both the query and database images are geo-tagged, and the location information is employed to verify whether the retrieval result is correct. If the distance between a candidate image and query is less than 25 meters, the candidate image will be viewed as a correct result. We use the popular evaluation metric recall@top$N$ to evaluate the performance, which regards the query is successfully recognized if there exist positive images among the top N results.

2) SAN FRANCISCO LANDMARK DATASET [6]

The dataset consists of 803 query images captured by mobile phones. The database contains 1.06M images collected by mobile mapping cars cruising in San Francisco city. Query images of the San Francisco landmark dataset are captured in the daytime, so the intensity filtering step will not be performed on this dataset. Each database image is annotated with longitude and latitude coordinates which can be used to discover the spatial relation of initial results. The ''carto id'' tag of database images is utilized for evaluation. A query image is regarded as correctly recognized if the top retrieved results share the same ''carto id'' as the query.

3) NetVLAD

For the NetVLAD feature, we use the public code and model released in the project homepage of [12]. The best NetVLAD model with intra-whiten trained on the Tokyo TM dataset is adopted for the Tokyo 24/7 dataset in our experiments. For the San Francisco landmark dataset, we use the NetVLAD model trained on the Pittsburgh dataset as its database images are very similar to the San Francisco dataset. This may be different from the result reported in the original paper, but has slight affect on the final result. For each query image, we adjust it to a maximum size of 640×320 pixels and keep the height-width ratio.

B. PARAMETERS ANALYSIS

For the intensity filtering step, the only parameter is the τ to distinguish dark images from bright images. Based on the test in Section III, we conduct an experiment to select the best parameter. We evaluate different thresholds to filter retrieval images with a low intensity, as illustrated in Fig. 8. The intensity filtering can improve the recall in a range of threshold, and we set τ to 45 in the following experiments. In details, the top-5 recall of the original NetVLAD is 81.9%, and the recall after intensity filtering is 84.4%. The intensity filtering improve the recall with a small margin on the Tokyo 24/7 dataset, but which is very useful for the group fusion step as the dark background images are removed from the top-ranked result list.

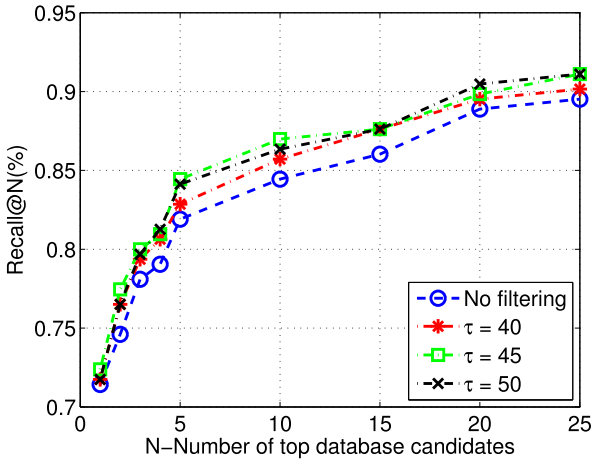

FIGURE 8. Intensity parameter selection.

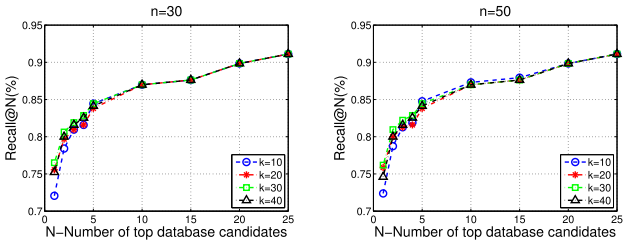

FIGURE 9. Group Fusion parameter selection on the Tokyo 24/7 dataset.

The intensity filtering can be employed to filter irrelevant results that most of their pixels are dark, such as the example of the first query in Fig. 1. However, for images with part dark background such as the example in Fig. 5, this schema fails. Our aim is to remove some obvious irrelevant results to facilitate the following group fusion of places. The problem of distinguishing images with dark background from image with bright background in fine-grained can be further explored in the future work.

In the group fusion step, we should set the rerank size n and the neighborhood parameter k. The rerank size n represents the group fusion should be performed on the top n initial results. The neighborhood parameter k is used to

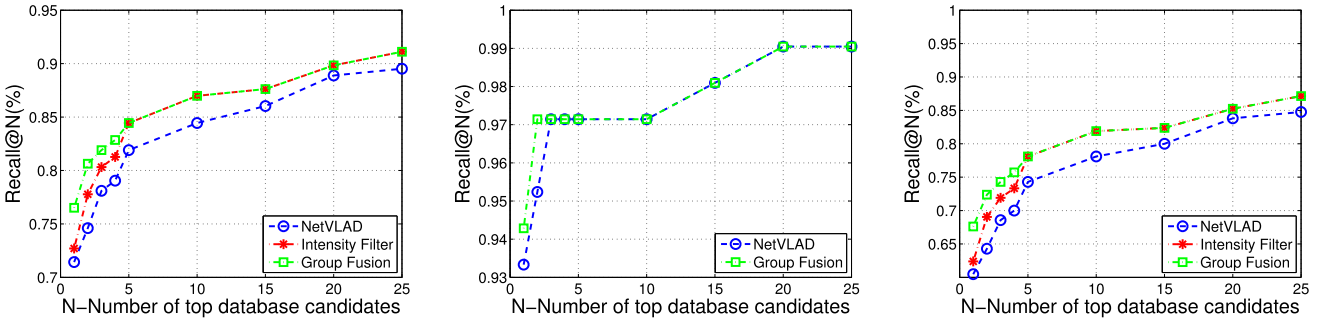

FIGURE 10. Illustration of the result of the proposed approach on the Tokyo 24/7 dataset. The intensity filtering is performed only for queries captured at sunset and night. (a) Tokyo 24/7 all queries. (b) Tokyo 24/7 daytime. (c) Tokyo 24/7 sunset/night.

determine the reciprocal neighborhood relationship. To avoid performance degeneration, we restrict n less than 50, which is also similar to the experimental setting of place recognition in [23]. We test different settings of n and k, and the result is illustrated in Fig. 9. As we only rerank the most relevant result at top-1, the curve of recall@N will be overlap for different parameters when the size of candidate database images N is large. For n is 30 or 50, the recall@1 will achieve a performance peak at $k = 30$, which implies a reliable setting to represent reciprocal neighborhood relationship. We set both k and n to 30 in the following experiments.

For the San Francisco landmark dataset, we also set both the k and n to 30 based on the similar parameter analysis process. As the dataset contains 1.06M database images, the computation of finding nearest neighborhoods for all database images is time cost. To reduce the offline computation time, we only consider the top-1000 initial results to build the reciprocal graph.

C. COMPARISON WITH STATE-OF-THE-ART

The proposed intensity filtering and group fusion of places are employed to rerank the initial results. We compare the proposed approach with the state-of-the-art approach NetVLAD [12]. Based on the experiment of parameter selection, the reranking is only performed on top-30 ranked results to bring the best result to the top and avoid degenerating performance. Similar to [11] and [12], we report the result of all queries, queries at day time and queries at sunset and night respectively on the Tokyo 24/7 dataset. For the San Francisco landmark dataset, we directly report the result of all queries.

As we can see from Fig. 10, the proposed approach can improve the recognition recall over the current state-of-the-art approach [12] on the Tokyo 24/7 dataset, which means relevant results are correctly reranked to the top. In details, the recall@top1 is improved from 71.4% to 76.5%, and the overall recall is also improved. The recall improvement before top-5 can be attributed to the group fusion step, and the other recall improvement can be attributed to the intensity filtering step. The main improvement of the proposed approach is from queries captured at sunset and night, and the intensity filtering strategy can filter some obvious irrelevant results effectively. The group fusion of places can improve recall@1 for queries captured at daytime, sunset and night.

For queries captured at daytime, the intensity filtering step will hurt the performance. Thus, the intensity filtering will not be performed for queries captured at daytime. As there exist rich details on these images, the powerful NetVLAD feature [12] can easily retrieve relevant results. With the group fusion step, the initial result list of queries at daytime can also be refined. Some successful examples by the proposed approach are illustrated in Fig. 12.

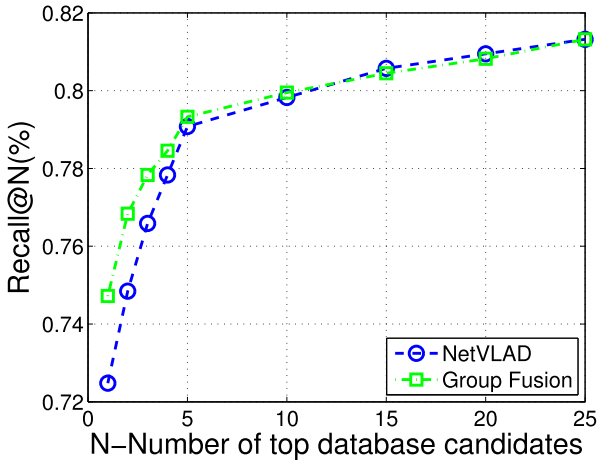

FIGURE 11. Illustration of the result of the proposed approach on the San Francisco landmark dataset. Please note that we only report the result after the graph fusion step. The intensity filtering will not be performed on this dataset as queries are captured in daytime.

For the San Francisco landmark dataset, the evaluated result is illustrated in Fig. 11. The group fusion step can select relevant results for the query and boost the recognition recall of top results. For example, the recall@top1 is improved from 72.4% to 74.7%, i.e., the proposed approach can find relevant result and bring it to the top. Please note that the proposed group fusion only considers the top-30 initial results returned by the NetVLAD based retrieval approach and reranks the most relevant result to the top-1 position.

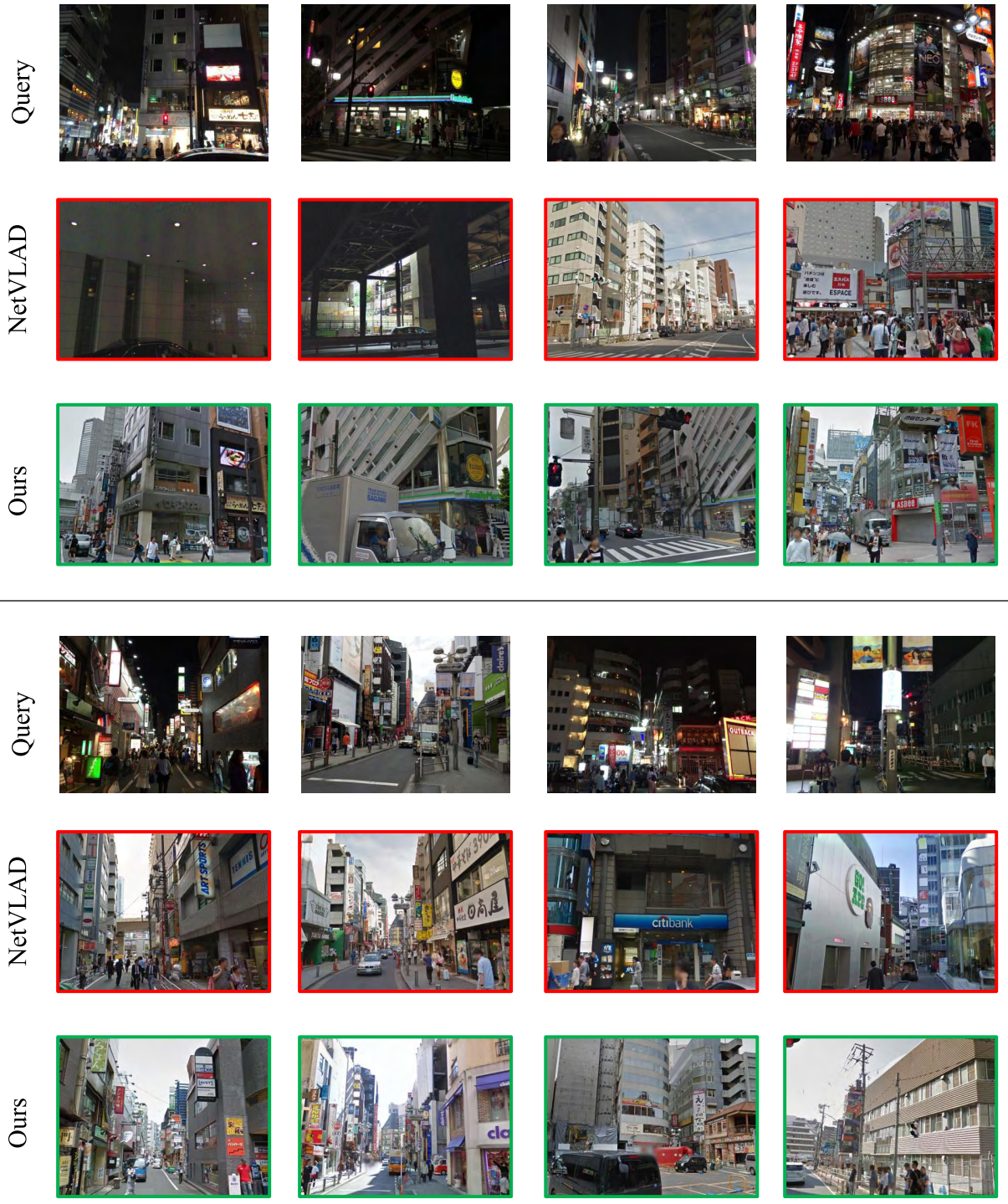

FIGURE 12. Some query examples by different approaches on the Tokyo 24/7 dataset. For each column, the first image represents the query, and the second and third images represent the top-1 query result by NetVLAD and the proposed approach respectively. Red rectangles mean irrelevant results, and green rectangles mean relevant results. The proposed approach can rerank relevant images to the top.

Therefore, the recognition result will be the same to the NetVLAD based approach as the size of top candidates increases.

In summary, the intensity filtering can remove top-ranked dark background results in the Tokyo 24/7 dataset, and the group fusion step can bring relevant results to the top for both datasets. By considering the correlation between top-ranked results, we can improve the recognition recall of top results. Typically, the intensity value and top nearest neighborhood results of database image can be calculated offline. The extra time cost of intensity filtering and graph fusion is low as the rerank process only considers initial top-ranked results.

VI. CONCLUSION

In this paper, we concentrate on improving the performance of NetVLAD feature for mobile place recognition. A filtering strategy is proposed to remove irrelevant results in coarse grain by verifying the intensity of images. Then, the adjacent groups are determined based on the spatial location of initial top-ranked results. By exploiting the reciprocal neighborhood relationship of initial results, we can construct the reciprocal neighborhood graph and discover the underlying correlation between initial results. Based on the graph, the initial results are evaluated to find the most possible relevant result of the query and bring it to the top. Experimental results on the Tokyo 24/7 and San Francisco landmark datasets demonstrate the effectiveness of the proposed approach.

As the NetVLAD is trained by using the Google Street View data collected at daytime, there still exist some hard queries that cannot retrieve any relevant images for the Tokyo 24/7 dataset. However, we have found that the Dense VLAD feature can be utilized as a complement when the NetVLAD feature fails. The following work will focus on fusing the result by different features or approaches for further improvement.

REFERENCES

[1] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 612–622, Apr. 2014.

[2] T. Guan, Y. He, L. Duan, J. Yang, J. Gao, and J. Yu, "Efficient BOF generation and compression for on-device mobile visual location recognition," *IEEE Multimedia*, vol. 21, no. 2, pp. 32–41, Apr./Jun. 2014.

[3] S. Gammeter, A. Gassmann, L. Bossard, T. Quack, and L. Van Gool, "Server-side object recognition and client-side object tracking for mobile augmented reality," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 1–8.

[4] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-DOF localization on mobile devices," in *Computer Vision*. Cham, Switzerland: Springer, 2014, pp. 268–283. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10605-2_18

[5] S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "I know what you did last summer: Object-level auto-annotation of holiday snaps," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 614–621.

[6] D. M. Chen *et al.*, "City-scale landmark identification on mobile devices," in *Proc. CVPR*, Jun. 2011, pp. 737–744.

[7] T. Sattler, B. Leibe, and L. Kobbelt, "Efficient & effective prioritized matching for large-scale image-based localization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1744–1756, Sep. 2017.

[8] M. Milford *et al.*, "Condition-invariant, top-down visual place recognition," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2014, pp. 5571–5577.

[9] N. Sünderhauf *et al.*, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Proc. Robot., Sci. Syst.*, Rome, Italy, Jul. 2015, pp. 199–208.

[10] X. Hu *et al.*, "Emotion-aware cognitive system in multi-channel cognitive radio ad hoc networks," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 180–187, Apr. 2018.

[11] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *Proc. CVPR*, Jun. 2015, pp. 1808–1817.

[12] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5297–5307.

[13] J. Sivic and A. Zisserman, "Video Google: Efficient visual search of videos," in *Toward Category-Level Object Recognition* (Lecture Notes in Computer Science), vol. 4170, J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, Eds. Heidelberg, Germany: Springer, 2006, pp. 127–144.

[14] R. Arandjelović and A. Zisserman, "DisLocation: Scalable descriptor distinctiveness for location recognition," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 188–204.

[15] G. Tolias, Y. Avrithis, and H. Jégou, "Image search with selective match kernels: Aggregation across single and multiple images," *Int. J. Comput. Vis.*, vol. 116, no. 3, pp. 247–261, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11263-015-0810-4

[16] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proc. CVPR*, Jun. 2013, pp. 883–890.

[17] M. Wang, L. Zhao, Y. Ming, E. Zhu, and J. Yin, "Boosting landmark retrieval baseline with burstiness detection," *IET Comput. Vis.*, vol. 12, no. 3, pp. 312–321, 2017.

[18] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys, "Large-scale location recognition and the geometric burstiness problem," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1582–1590.

[19] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2007, pp. 1–8.

[20] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit. (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2011, pp. 889–896.

[21] M. Wang, Y. Ming, Q. Liu, and J. Yin, "Fusion of global and local deep representation for effective object retrieval," in *Theoretical Computer Science*, D. Du, L. Li, E. Zhu, and K. He, Eds. Singapore: Springer, 2017, pp. 31–45.

[22] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012. [Online]. Available: https://hal.inria.fr/inria-00633013

[23] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific rank fusion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 803–815, Apr. 2015.

[24] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, Nov. 2004.

[25] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Miami, FL, USA: IEEE Computer Society, Jun. 2009, pp. 1169–1176. [Online]. Available: https://hal.inria.fr/inria-00394211

[26] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[27] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1011139631724

[28] S. Cao and N. Snavely, "Graph-based discriminative learning for location recognition," *Int. J. Comput. Vis.*, vol. 112, no. 2, pp. 239–254, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-014-0774-9

[29] D. Qin, S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Proc. CVPR*, Jun. 2011, pp. 777–784.

[30] H. Zhou, T. Sattler, and D. W. Jacobs, "Evaluating local features for day-night matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands, Oct. 2016, pp. 724–736.

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[32] J. L. Schönberger, T. Price, T. Sattler, J.-M. Frahm, and M. Pollefeys, "A vote-and-verify strategy for fast spatial verification in image retrieval," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2016, pp. 321–337.

[33] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2006.

[34] M. Shi, Y. Avrithis, and H. Jégou, "Early burst detection for memory-efficient image retrieval," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 605–613.

[35] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Spatially-constrained similarity measurefor large-scale object retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1229–1241, Jun. 2014.

MAO WANG received the B.S. degree in computer science and technology from the National University of Defense Technology (NUDT), China, in 2011, and the M.S. degree in computer science and technology from the Beijing Institute of Technology, China, in 2014. He is currently pursuing the Ph.D. degree with College of Computer, NUDT. His research interests include image retrieval and machine learning.

EN ZHU received the M.S. and Ph.D. degrees in computer science from the National University of Defense Technology, Changsha, China, in 2001 and 2005, respectively. From 2009 to 2010, he visited the Department of Computer Science, University of York, York, U.K. He is currently with the School of Computer, National University of Defense Technology. His main research interests are pattern recognition, image processing, and machine learning.

QIANG LIU (M'14) received the Ph.D. degree in computer science and technology from the National University of Defense Technology (NUDT) in 2014. From 2011 to 2013, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, The University of British Columbia, Canada. He is currently an Assistant Professor with NUDT. His research interests include 5G network, Internet of Things, wireless network security, and machine learning. He is a member of the China Computer Federation. He has contributed several archived journal and international conference papers, such as the *IEEE Network Magazine*, the IEEE Transactions on Wireless Communications, the IEEE Transactions on Cybernetics, the *Pattern Recognition*, the IEEE Communication Letters, the *Neurocomputing*, the *Neural Computation and Applications*, the *Mobile Information Systems*, EDBT'17, WCNC'17, ICANN'17, and SmartMM'17. He currently serves on the Editorial Review Board for the *Artificial Intelligence Research Journal*. He has served as the Co-Chair for SmartMM'18.

YONGKAI YE received the B.S. and M.S. degrees in computer science and technology from the National University of Defense Technology, Changsha, China, in 2010 and 2013, respectively, where he is currently pursuing the Ph.D. degree with the College of Computer. His research interests include machine learning and multikernel clustering.

YUEWEI MING received the B.S. degree from Sichuan University in 2011 and the M.S. degree from the National University of Defense Technology in 2013, where he is currently pursuing the Ph.D. degree. His research interests include distributed and parallel optimization and scalable machine learning systems.

JIANPING YIN received the M.S. and Ph.D. degrees in computer science from the National University of Defense Technology, Changsha, China, in 1986 and 1990, respectively. He is currently a Distinguished Professor of computer science with the Dongguan University of Technology. His research interests involve artificial intelligence, pattern recognition, algorithm design, and information security.