Avoiding confusing features in place recognition

Jan Knopp[1], Josef Sivic[2], Tomas Pajdla[3]

[1]VISICS, ESAT-PSI, K.U. Leuven, Belgium
[2]INRIA, WILLOW, Laboratoire d'Informatique de l'Ecole Normale Superieure, Paris*
[3]Center for Machine Perception, Czech Technical University in Prague

Abstract. We seek to recognize the place depicted in a query image using a database of "street side" images annotated with geolocation information. This is a challenging task due to changes in scale, viewpoint and lighting between the query and the images in the database. One of the key problems in place recognition is the presence of objects such as trees or road markings, which frequently occur in the database and hence cause significant confusion between different places. As the main contribution, we show how to avoid features leading to *confusion* of particular places by using geotags attached to database images as a form of supervision. We develop a method for automatic detection of image-specific and spatially-localized groups of confusing features, and demonstrate that suppressing them significantly improves place recognition performance while reducing the database size. We show the method combines well with the state of the art bag-of-features model including query expansion, and demonstrate place recognition that generalizes over wide range of viewpoints and lighting conditions. Results are shown on a geotagged database of over 17K images of Paris downloaded from Google Street View.

1 Introduction

Map-based collections of street side imagery, such as Google StreetView [1] or Microsoft StreetSide [2] open-up the possibility of image-based place recognition. Given the query image of a particular street or a building facade, the objective is to find one or more images in the geotagged database *depicting the same place*. We define "place" as the 3D structure visible in the query image, rather than the actual camera location of the query [3]. Images showing (a part of) the same 3D structure may, and often have, very different camera locations, as illustrated in the middle column of figure 1.

The ability to visually recognize the place depicted in an image has a range of exciting applications such as: (i) automatic registration of consumer photographs

* CNRS/ENS/INRIA UMR 8548

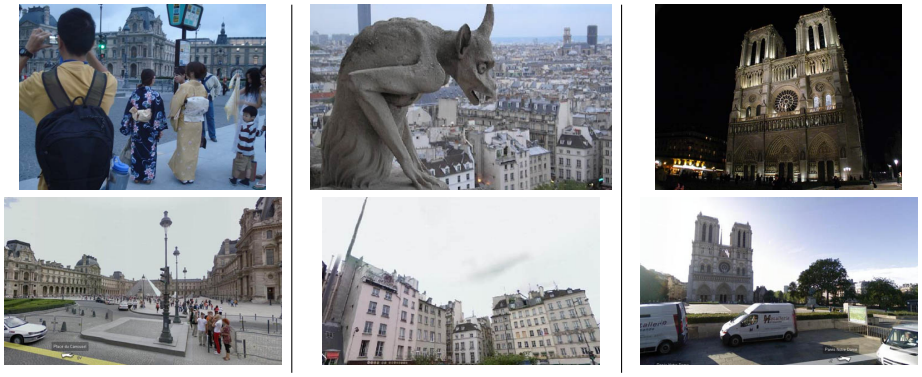


Fig. 1. Examples of visual place recognition results. Given a query image (top) of an unknown place, the goal is to find an image from a geotagged database of street side imagery (bottom), depicting the same place as the query.

with maps [4], (ii) transferring place-specific annotations, such as landmark information, to the query image [5, 6], or (iii) finding common structures between images for large scale 3D reconstruction [7]. In addition, it is an important first step towards estimating the actual query image camera location using structure from motion techniques [8, 9, 7].

Place recognition is an extremely challenging task as the query image and images available in the database might show the same place imaged at a different scale, from a different viewpoint or under different illumination conditions. An additional key challenge is the self-similarity of images of different places: the image database may contain objects, such as trees, road markings or window blinds, which occur at many places and hence are not representative for any particular place. In turn, such objects significantly confuse the recognition process.

As the main contribution of this work, we develop a method for automatically detecting such "confusing objects" and demonstrate that removing them from the database can significantly improve the place recognition performance. To achieve this, we employ the efficient bag-of-visual-words [10, 11] approach with large vocabularies and fast spatial matching, previously used for object retrieval in large unstructured image collections [12, 13]. However, in contrast to generic object retrieval, the place recognition database is structured: images depict a consistent 3D world and are labelled with geolocation information. We take advantage of this additional information and use the available geotags as a form of supervision providing us with large amounts of negative training data since images from far away locations cannot depict the same place. In particular, we detect, in each database image, spatially localized groups of local invariant features, which are matched to images far from the geospatial location of the database image. The result is a segmentation of each image into a "confusing layer", represented by groups of spatially localized invariant features occurring at other places in the database, and a layer discriminating the particular place

from other places in the database. Further, we demonstrate that suppressing such confusing features significantly improves place recognition performance while reducing the database size.

To achieve successful visual place recognition the image database has to be representative: (i) all places need to be covered and (ii) each place should be captured under wide range of imaging conditions. For this purpose we combine two types of visual data: (i) street-side imagery from Google street-view which has good coverage and provides accurate geo-locations; and (ii) user-generated imagery from a photo-sharing website Panoramio, which depicts places under varying imaging conditions (such as different times of the day or different seasons), but is biased towards popular places and its geotags are typically noisy. We show place recognition results on a challenging database of 17K images of central Paris automatically downloaded from Google Street View expanded with 8K images from the photo-sharing website Panoramio.

1.1 Related work

Most previous work on image-based place recognition focused on small scale settings [14–16]. More recently, Cummins and Newman [17] described an appearance-only simultaneous localization and mapping (SLAM) system, based on the bag-of-features representation, capturing correlations between different visual words. They show place recognition results on a dataset of more than 100,000 omni-directional images captured along a 1,000 km route, but do not attempt to detect or remove confusing features. Schindler *et al.* [3] proposed an information theoretic criterion for choosing informative features for each location, and build vocabulary trees [18] for location recognition in a database of 30,000 images. However, their approach relies on significant visual overlap between spatially close-by database images, effectively providing positive "training data" for each location. In contrast, our method measures only statistics of mismatched features and requires only negative training data in the form of highly ranked mismatched images for a particular location.

Large databases of several millions of geotagged Flickr images were recently used for coarse-level image localization. Hays and Efros [19] achieve coarse-level localization on the level of continents and cities using category-level scene matching. Li *et al.* [6] discover distinct but coarse-level landmarks (such as an entire city square) as places with high concentration of geotagged Flickr images and build image-level classifiers to distinguish landmarks from each other. In contrast, we address the complementary task of matching particular places in street-side imagery, use multi-view spatial constraints and require establishing visual correspondence between the query and the database image.

Community photo-collections (such as Flickr) are now often used in computer vision tasks with the focus on clustering [20, 21, 5], 3D modelling [9, 7] and summarization [22]. In contrast, we combine images from a community photo-collection with street-side imagery to improve place recognition performance.

The task of place recognition is similar to object retrieval from large unstructured image collections [20, 23, 18, 13, 24], and we build on this work. However,

we propose to detect and suppress confusing features taking a strong advantage of the structured nature of the geolocalized street side imagery.

Finally, the task of confuser detection has some similarities with the task of feature selection in category-level recognition [25–27] and retrieval [28–30]. These methods typically learn discriminative features from clean labelled data in the Caltech-101 like setup. We address the detection and suppression of spatially localized groups of confusing (rather than discriminative) features in the absence of positive (matched) training examples, which are not directly available in the geo-referenced image collection. In addition, we focus on matching particular places under viewpoint and lighting variations, and in a significant amount of background clutter.

The reminder of the paper is organized as follows. Section 2 reviews the baseline place recognition algorithm based on state-of-the-art bag-of-features object retrieval techniques. In section 3 we describe the proposed method for detection of spatially localized groups of confusing features and in section 4 we outline how the detected confusers are avoided in large scale place matching. Finally, section 5 describes the collected place recognition datasets and experimentally evaluates the benefits of suppressing confusers.

2 Baseline place recognition with geometric verification

We have implemented a two-stage place recognition approach based on state-of-the-art techniques used in large scale image and object retrieval [18, 13]. In the first stage, the goal is to efficiently find a small set of candidate images (50) from the entire geotagged database, which are likely to depict the correct place. This is achieved by employing the bag-of-visual-words image representation and fast matching techniques based on inverted file indexing. In the second verification stage, the candidate images are re-ranked taking into account the spatial layout of local quantized image features. In the following we describe our image representation and give details of the implementation of the two image matching stages.

Image representation: We extract SURF [31] features from each image. They are fast to extract (under one second per image), and we have found them to perform well for place recognition in comparison with affine invariant features frequently used for large-scale image retrieval [23, 18, 13] (experiments not shown in the paper). The extracted features are then quantized into a vocabulary of 100K visual words. The vocabulary is built from a subset of 2942 images (about 6M features) of the geotagged image database using the approximate k-means algorithm [32, 13]. Note that as opposed to image retrieval, where generic vocabularies trained from a separate training dataset have been recently used [23], in the context of location recognition a vocabulary can be trained for a particular set of locations, such as a district in a city.

Initial retrieval of candidate places: Similar to [13], both the query and database images are represented using tf-idf [33] weighted visual word vectors and the

similarity between the query and each database vector is measured using the normalized scalar product. The tf-idf weights are estimated from the entire geotagged database. This type of image matching has been shown to perform near real-time matching in datasets of 1M images [23, 18, 13]. After this initial retrieval stage we retain the top 50 images ranked by the similarity score.

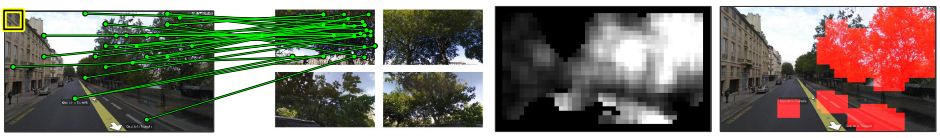
Filtering by spatial verification: In the second stage we filter the candidate set using a test on consistency of spatial layout of local image features. We assume that the 3D structure visible in the query image and each candidate image can be approximated by a small number of planes (1-5) and fit multiple homographies using RANSAC with local optimization [34]. The piecewise planar approximation has the benefit of increased efficiency and has been shown to perform well for matching in urban environments [13]. The candidate images are then re-ranked based on the number of inliers.

Enhancing street-side imagery with additional photographs: In image retrieval query expansion has been shown to significantly improve retrieval performance by enhancing the original query using visual words from spatially-verified images in the database [12]. Here, we perform query expansion using a collection of images downloaded from a photo-sharing site and details of this data will be given in section 5. These images are not necessarily geotagged, but might contain multiple images of the same places captured by different photographers from different viewpoints or different lighting conditions. The place recognition algorithm then proceeds in two steps. First the query image is expanded by matching to the non-geotagged database. Second, the enhanced query image is used for the place recognition query to the geotagged database. We implement the "average query expansion" described in [12].

3 Detecting spatially localized groups of confusing features

Locations in city-street image databases contain significant amount of features on objects like trees or road markings, which are not informative for recognizing a particular place since they appear frequently throughout the city. This is an important problem as such features pollute the visual word vectors and can cause significant confusion between different places. To address this issue we focus in this section on automatically detecting such regions. To achieve this, we use the fact that *an image of a particular place should not match well to other images at far away locations*. The details of the approach are given next.

Local confusion score: For each database image I, we first find a set $\{I_n\}$ of top n "confusing" images from the geotagged database. This is achieved by retrieving top matching images using fast bag-of-visual-words matching (section 2), but excluding images at locations closer than d_{min} meters from the location of I to ensure that retrieved images do not contain the same scene. A local confusion



(a) (b) (c)

Fig. 2. Detection of place-specific confusing regions. (a) Features in each database image are matched with features of similar images at geospatially far away locations (illustration of matches to only one image is shown). (b) Confusion score is computed in a sliding window manner, locally counting the proportion of mismatched features. Brightness indicates high confusion. (c) An image is segmented into a "confusing layer" (indicated by red overlay), and a layer (the rest of the image) discriminating the particular place from other places in the database.

score ρ is then measured over the image I in a sliding window manner on a dense grid of locations. For a window w at a particular image position we determine the score as

$$\rho_w = \sum_{k=1}^{n} \frac{M_w^k}{N_w}, \tag{1}$$

where M_w^k is the number of tentative feature matches between the window w and the k-th "confusing" image, and N_w is the total number of visual words within the window w. In other words, the score measures the number of image matches normalized by the number of detected features in the window. The score is high if a large proportion of visual words (within the window) matches to the set of confusing images and is low in areas with relatively small number of confusing matches. The confusion score can then be used to obtain a segmentation of the image into a layer specific for the particular place (regions with low confusion score) and a confuser layer (regions with high confusion score). In this work we opt for a simple threshold based segmentation, however more advanced segmentation methods respecting image boundaries can be used [35]. In addition, for a window to be deemed confusing, we require that $N_w > 20$, which ensures windows with a small number of feature detections (and often less reliable confusion score estimates) are not considered. The entire process is illustrated in figure 2. Several examples are shown in figure 3. The main parameters of the method are the width s of the sliding window and the threshold t on the confusion score. We set $s = 75$ pixels, where the windows are spaced on a 5 pixel grid in the image, and $t = 1.5$, i.e. a window has to have 1.5 times more matches than detected features to be deemed confusing. Sensitivity of the place recognition performance to selection of these parameters is evaluated in section 5.

4 Place matching with confuser suppression

The local confusion score can potentially be used in all stages of the place recognition pipeline, i.e., for vocabulary building, initial retrieval, spatial verification

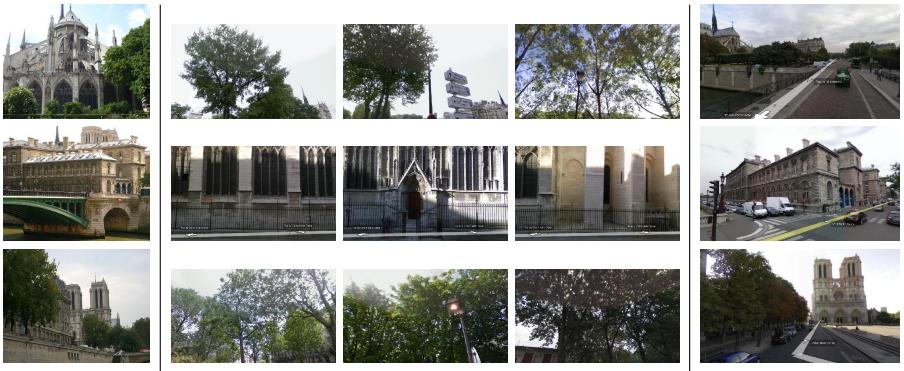
(a) (b) (c) (d)

Fig. 3. Examples of detected confusing regions which are obtained by finding local features in original image (a) frequently mismatched to similar images of different places shown in (b). (c) Detected confusing image regions. (d) Features within the confusing regions are erased (red) and the rest of features are kept (green). Note that confusing regions are spatially localized and fairly well correspond to real-world objects, such as trees, road, bus or a window blind. Note also the different geospatial scale of the detected "confusing objects": trees or pavement (top two rows) might appear anywhere in the world; a particular type of window blinds (3rd row) might be common only in France; and the shown type of bus (bottom row) might appear only in Paris streets. Confusing features are also place specific: trees deemed confusing at one place, might not be detected as confusing at another place, depending on the content of the rest of the image. Note also that confusion score depends on the number of detected features. Regions with no features, such as sky, are not detected.

and query expansion. In the following we investigate suppressing confusers in the initial retrieval stage.

To understand the effect of confusers on the retrieval similarity score $s(\mathbf{q}, \mathbf{v}^i)$ between the query $\mathbf{q}$ and each database visual word vector $\mathbf{v}^i$ we can write both the query and the database vector as $\mathbf{x} = \mathbf{x}_p + \mathbf{x}_c$, where $\mathbf{x}_p$ is place specific and $\mathbf{x}_c$ is due to confusers. The retrieval score is measured by the normalized scalar product (section 2),

$$s(\mathbf{q}, \mathbf{v}^i) = \frac{\mathbf{q}^\top \mathbf{v}^i}{\|\mathbf{q}\|\|\mathbf{v}^i\|} = \frac{(\mathbf{q}_p + \mathbf{q}_c)^\top (\mathbf{v}_p^i + \mathbf{v}_c^i)}{\|\mathbf{q}_p + \mathbf{q}_c\|\|\mathbf{v}_p^i + \mathbf{v}_c^i\|} = \frac{\mathbf{q}_p{}^\top \mathbf{v}_p^i + \mathbf{q}_c{}^\top \mathbf{v}_p^i + \mathbf{q}_p{}^\top \mathbf{v}_c^i + \mathbf{q}_c{}^\top \mathbf{v}_c^i}{\|\mathbf{q}_p + \mathbf{q}_c\|\|\mathbf{v}_p^i + \mathbf{v}_c^i\|}. \quad (2)$$



(a) (b) (c)

Fig. 4. Improvement in place recognition based on suppressing confusing features. (a) The query image. (b) Three top ranked images after initial retrieval and spatial verification. (c) The top ranked image after suppressing confusing image regions. Note that the highly ranked false positive images shown in (b) are suppressed in (c).

If confusers are detected and removed in each database image the terms involving $\mathbf{v}_c^i$ vanish. Further, if there are no common features between $\mathbf{q}_c$ and $\mathbf{v}_p^i$, i.e. confusers in the query image do not intersect with place specific features in the database, ${\mathbf{q}_c}^\top \mathbf{v}_p^i = 0$. Under these two assumptions, the retrieval score reduces to

$$s(\mathbf{q}, \mathbf{v}^i) = \frac{1}{\|\mathbf{q}_p + \mathbf{q}_c\|} \frac{1}{\|\mathbf{v}_p^i\|} ({\mathbf{q}_p}^\top \mathbf{v}_p^i) \propto \frac{1}{\|\mathbf{v}_p^i\|} ({\mathbf{q}_p}^\top \mathbf{v}_p^i) \tag{3}$$

For a given query, we are interested only in the ranking of the database images and not the actual value of the score, hence the query normalization dependent on $\mathbf{q}_c$ can be ignored. This is an interesting property as it suggests that if all confusers are removed from the database, the ranking of database images does not depend on confusers in the query. In practice, however, the second assumption above, ${\mathbf{q}_c}^\top \mathbf{v}_p^i = 0$, might not be always satisfied, since confusers are specific to each place, and not necessary global across the whole database. Hence, some common features between $\mathbf{q}_c$ and $\mathbf{v}_p^i$ may remain. Nevertheless, we demonstrate significant improvements in place recognition (section 5) by suppressing confusers on the database side, i.e. setting $\mathbf{v}_c^i = \mathbf{0}$ for all database images and implicitly exploiting the fact that ${\mathbf{q}_c}^\top \mathbf{v}_p^i \ll {\mathbf{q}_p}^\top \mathbf{v}_p^i$.

Implementation: The local confusion score is pre-computed offline for each image in the database, and all features with a score greater than a certain threshold are suppressed. The remaining features are then indexed using visual words. The initial retrieval, spatial verification and query expansion are performed as outlined in section 2 but for initial retrieval we remove confusing features from the geotagged database. The benefits of suppressing confusing features for place recognition are illustrated in figure 4.

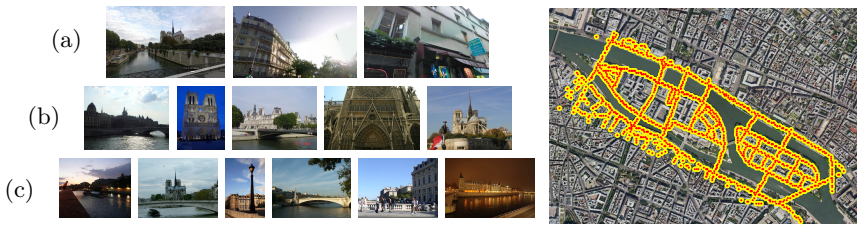


Fig. 5. Left: examples of (a) geo-tagged images; (b) test query images; (c) non-geotagged images. Right: locations of geo-tagged images overlaid on a map of Paris.

Discussion: Note that the proposed confusion score is different from the tf-idf weighting [33], typically used in image retrieval [36, 18, 13, 24], which down-weights frequently occurring visual words in the whole database. The tf-idf score is computed independently for each visual word and estimated globally based on the frequency of occurrence of visual words in the whole database, whereas in our case the confusion score is estimated for a local window in each image. The local confusion score allows removing confusers that are specific to particular images and avoids excessive pruning of features that are confusing in some but hold useful information for other images. Moreover, the top-retrieved images from faraway places, which are used to determine the confusion score, act as place-specific difficult negative "training" examples. This form of supervision is naturally available for georeferenced imagery, but not in the general image retrieval setting. This type of negative supervisory signal is also different from the clean (positive and negative) supervision typically used in feature selection methods in object category recognition [25–27] and retrieval [28–30]. In our case, obtaining verified positive examples would require expensive image matching, and for many places positive examples are not available due to sparse location sampling of the image database.

5 Experimental evaluation

First, we describe image datasets and the performance measure, which will be used to evaluate the proposed place recognition method. In the following subsection, we test the sensitivity to key parameters and present place recognition results after different stages of the algorithm.

5.1 Image datasets

Geotagged google street-view images: The geotagged dataset consists of about 17K images automatically downloaded from Google StreetView [1]. We have downloaded all available images in a district of Paris covering roughly an area of 1.7×0.5 kilometers. The full 360×180 panorama available at each distinct location is represented by 12 perspective images with resolution 936×537 pixels. Example images are shown in figure 5(a) and image locations overlaid on a map are shown in figure 5(right).

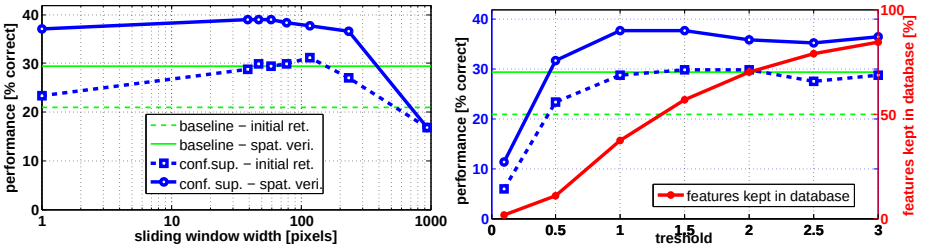
Non-geotagged images: Using keyword and location search we have downloaded about 8K images from the photo-sharing website Panoramio [37]. Images were downloaded from roughly the same area as covered by the geotagged database. The location information on photo-sharing websites is very coarse and noisy and therefore some images are from other parts of Paris or even different cities. Apart from choosing which images to download, we do not use the location information in any stage of our algorithm and treat the images as non-geotagged.

Test set: In addition, a test set of 200 images was randomly sampled from the non-geotagged image data. These images are set aside as unseen query images and are not used in any stage of the processing apart from testing. Examples of query images and non-geotagged images are shown in figure 5 (b) and (c).

Performance measures: Given a test query image the goal is to recognize the place by finding an image from the geotagged database depicting the same place, i.e., the same 3D structure. We measure the recognition performance by the number of test images (out of 200 test queries), for which the top-ranked image from the geotagged database correctly depicts the same place. The ground truth is obtained manually by inspection of the visual correspondence between the query and the top retrieved image. The overall performance is then measured by the percentage of correctly matched test images. As 33 images (out of the 200 randomly sampled queries) do not depict places within the geotagged database, the perfect score of 100% would be achieved when the remaining 167 images are correctly matched.

5.2 Performance evaluation

Parameter settings: We have found that parameter settings of the baseline place recognition, such as the vocabulary size K (=10^5), the top m (=50) candidates for spatial verification or the minimum number of inliers (20) to deem a successful match work well with confuser suppression and keep them unchanged throughout the experimental evaluation. For confuser suppression, we set the minimal spatial distance to obtain confusing images to one fifth of the map (about 370 meters) and consider the top $n = 20$ confusing images. In the following, we evaluate sensitivity of place recognition to the sliding window width, s, and confuser score threshold, t. We explore two one-dimensional slices of the 2-D parameter space, by varying s for fixed $t = 1.5$, figure 6(a)), and varying t for fixed $s =$ 75 pixels, (figure 6(b)). From graph 6(a), we note that a good performance is obtained for window sizes between 30 and 100 pixels. The window size specially affects the performance of the initial bag-of-visual-words matching and less so the results after spatial verification. This may be attributed to a certain level of spatial consistency implemented by the intermediate-size windows, where groups of spatially-localized confusing features are removed. However, even removing individual features (s=1 pixel) enables retrieving many images, initially low-ranked by the baseline approach, within the top 50 matches so that they are later



(a) (b)

Fig. 6. (a) Place recognition performance for varying confuser sliding window width s. (b) Place recognition performance (left axis) and percentage of features kept in the geotagged database (right axis) for varying confuser detection threshold t.

Method	**% correct** *initial retrieval*	**% correct** *with spatial verification*
a. Baseline place recognition	20.96	29.34
b. Query expansion	26.35	41.92
c. Confuser suppression	29.94	37.72
d. Confuser suppression+Query expansion	32.93	**47.90**

Table 1. Percentage of correctly localized test queries for different place recognition approaches.

correctly re-ranked with spatial verification. Graph 6(b) again shows good place recognition performance over a wide range of confuser detection thresholds. The chosen value $t = 1.5$ represents a good compromise between the database size and place recognition performance, keeping around 60% of originally detected features. However, with a small loss in initial retrieval performance, even a lower threshold $t = 1$ can be potentially used.

Overall place recognition performance: In the reminder of this section, we evaluate the overall place recognition performance after each stage of the proposed method. Results are summarized in table 1. It is clear that spatial re-ranking improves initial bag-of-visual-words matching in all stages of the proposed algorithm. This illustrates that the initial bag-of-visual words matching can be noisy and does not always return the correct match at the top rank, however, correct matches can be often found within the top 50 best matches. Both the query expansion and non-informative feature suppression also significantly improve place recognition performance of the baseline approach. When applied together, the improvement is even bigger correctly recognizing 47.90% of places in comparison with only 41.92% using query expansion alone and 37.72% using confuser suppression alone. This could be attributed to the complementarity of both methods. The place query expansion improves recall by enhancing the query using relevant features found in the non-geotagged database, whereas confuser suppression removes confusing features responsible for many highly ranked false

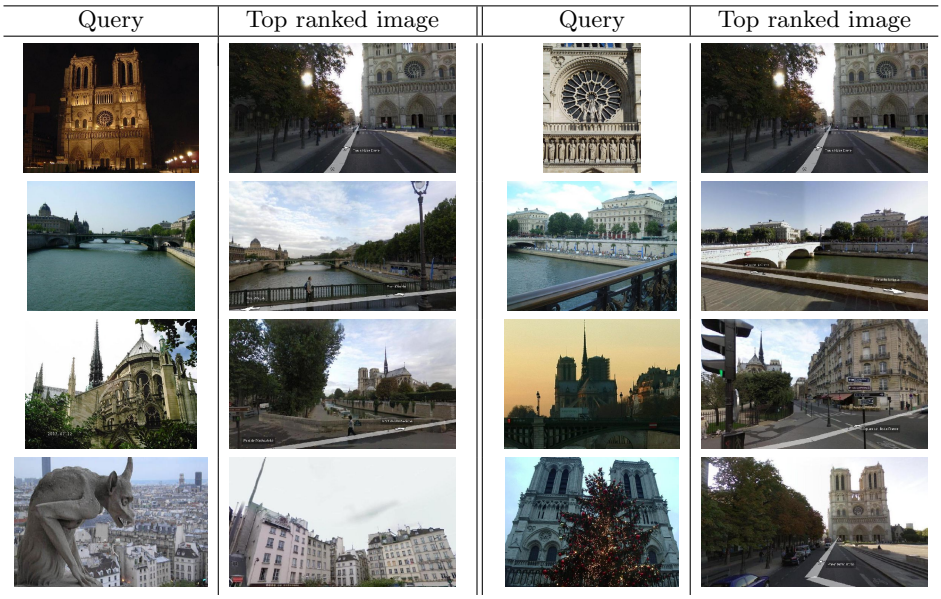


Fig. 7. Examples of correct place recognition results. Each image pair shows the query image (left) and the best match from the geotagged database (right). Note that query places are recognized despite significant changes in viewpoint (bottom left), lighting conditions (top left), or presence of large amounts of clutter and occlusion (bottom right).

Fig. 8. Examples of challenging test query images, which were not found in the geotagged database.

positives. Overall, the performance with respect to the baseline bag-of-visual-words method (without spatial re-ranking) is more than doubled from 20.96% to 47.90% correctly recognized place queries – a significant improvement on the challenging real-world test set. Examples of correct place recognition results are shown in figure 7. Examples of non-localized test queries are shown in figure 8. Many of the non-localized images represent very challenging examples for current matching methods due to large changes in viewpoint, scale and lighting conditions. It should be also noted that the success of query expansion depends on the availability of additional photos for a particular place. Places with additional images have a higher chance to be recognized.

6 Conclusions

We have demonstrated that place recognition performance for challenging real-world query images can be significantly improved by automatic detection and suppression of spatially localized groups of confusing non-informative features in the geotagged database. Confusing features are found by matching places spatially far on the map – a negative supervisory signal readily available in geotagged databases. We have also experimentally demonstrated that the method combines well with the state of the art bag-of-features model and query expansion.

Detection of spatially defined confusing image regions opens up the possibility of their automatic clustering and category-level analysis (when confusers correspond to trees, pavement or buses), determining their geospatial scale (trees might appear everywhere, whereas a particular type of buses may not), and reasoning about their occurrence in conjunction with location-specific objects (a tree in front of a house may still be a characteristic feature). Next, we plan to include such category-level place analysis in the current framework to further improve the place recognition performance.

Acknowledgements: We are grateful for financial support from the MSR-INRIA laboratory, ANR-07-BLAN-0331-01, FP7-SPACE-241523 PRoViScout and MSM6840770038.

References

1. (http://maps.google.com/help/maps/streetview/)
2. (http://www.bing.com/maps/)
3. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: CVPR. (2007)
4. Aguera y Arcas, B.: (Augmented reality using Bing maps.) Talk at TED 2010.
5. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: CIVR. (2008)
6. Li, Y., Crandall, D., Huttenlocher, D.: Landmark classification in large-scale image collections. In: ICCV. (2009)
7. Snavely, N., Seitz, S., Szeliski, R.: Photo tourism: exploring photo collections in 3D. In: SIGGRAPH. (2006)
8. Havlena, M., Torii, A., Pajdla, T.: Efficient structure from motion by graph optimization. In: ECCV. (2010)
9. Schaffalitzky, F., Zisserman, A.: Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In: ECCV. (2002)
10. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: WS-SLCV, ECCV. (2004)
11. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: ICCV. (2003)
12. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: ICCV. (2007)
13. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR. (2007)

14. Shao, H., Svoboda, T., Tuytelaars, T., van Gool, L.: Hpat indexing for fast object/scene recognition based on local appearance. In: CIVR. (2003)
15. Silpa-Anan, C., Hartley, R.: Localization using an image-map. In: ACRA. (2004)
16. Zhang, W., Kosecka, J.: Image based localization in urban environments. In: 3DPVT. (2006)
17. Cummins, M., Newman, P.: Highly scalable appearance-only SLAM - FAB-MAP 2.0. In: Proceedings of Robotics: Science and Systems, Seattle, USA (2009)
18. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR. (2006)
19. Hays, J., Efros, A.: im2gps: estimating geographic information from a single image. In: CVPR. (2008)
20. Chum, O., Perdoch, M., Matas, J.: Geometric min-hashing: Finding a (thick) needle in a haystack. In: CVPR. (2009)
21. Li, X., Wu, C., Zach, C., Lazebnik, S., J.-M., F.: Modeling and recognition of landmark image collections using iconic scene graphs. In: ECCV. (2008)
22. Simon, I., Snavely, N., Seitz, S.: Scene summarization for online image collections. In: SIGGRAPH. (2006)
23. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large-scale image search. In: ECCV. (2008)
24. Turcot, P., Lowe, D.: Better matching with fewer features: The selection of useful features in large database recognition problem. In: WS-LAVD, ICCV. (2009)
25. Lee, Y., Grauman, K.: Foreground focus: Unsupervised learning from partially matching images. IJCV **85** (2009)
26. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR. (2006)
27. Torralba, A., Murphy, K., Freeman, W.: Sharing visual features for multiclass and multiview object detection. IEEE PAMI **29** (2007)
28. Kulis, B., Jain, P., Grauman, K.: Fast similarity search for learned metrics. IEEE PAMI **31** (2009)
29. Torresani, L., Szummer, M., Fitzgibbon, A.: Learning query-dependent prefilters for scalable image retrieval. In: CVPR. (2009)
30. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: ICCV. (2007)
31. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: ECCV. (2006)
32. Muja, M., Lowe, D.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP. (2009)
33. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management **24** (1988)
34. Chum, O., Matas, J., Obdrzalek, S.: Enhancing RANSAC by generalized model optimization. In: ACCV. (2004)
35. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: ICCV. (2001)
36. Jegou, H., Douze, M., Schmid, C.: On the burstiness of visual elements. In: CVPR. (2009)
37. (http://www.panoramio.com/)