City-Scale Landmark Identification on Mobile Devices

David M. Chen [1] Georges Baatz [2] Kevin Köser [2] Sam S. Tsai [1]
Ramakrishna Vedantham [3] Timo Pylvänäinen [3] Kimmo Roimela [3] Xin Chen [4]
Jeff Bach [4] Marc Pollefeys [2] Bernd Girod [1] Radek Grzeszczuk [3]

[1] Stanford University [2] ETH Zurich [3] Nokia Research Center [4] NAVTEQ

Abstract

With recent advances in mobile computing, the demand for visual localization or landmark identification on mobile devices is gaining interest. We advance the state of the art in this area by fusing two popular representations of street-level image data—facade-aligned and viewpoint-aligned—and show that they contain complementary information that can be exploited to significantly improve the recall rates on the city scale. We also improve feature detection in low contrast parts of the street-level data, and discuss how to incorporate priors on a user's position (e.g. given by noisy GPS readings or network cells), which previous approaches often ignore. Finally, and maybe most importantly, we present our results according to a carefully designed, repeatable evaluation scheme and make publicly available a set of 1.7 *million images with ground truth labels, geotags, and calibration data, as well as a difficult set of cell phone query images. We provide these resources as a benchmark to facilitate further research in the area.*

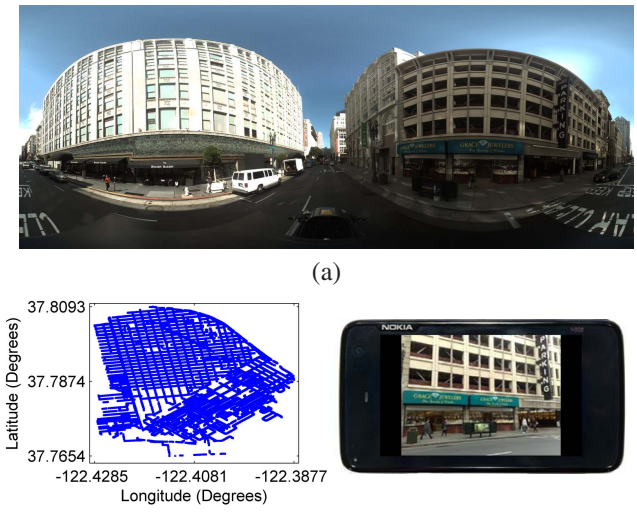


Figure 1. (a) Panorama from a large city-scale image database. (b) Locations in San Francisco where image data was collected. (c) Query image taken with a camera phone.

1. Introduction

This paper looks at the problem of city-scale landmark recognition from cell phone images. Research in this area has been motivated by emerging commercial applications for mobile devices like Google Goggles[1] and interest from the robotics community [4]. The main contributions of this work are twofold: we publish an extensive landmark dataset that will help to push research forward in the area of mobile location recognition, and we present a comprehensive location recognition pipeline that could serve as a baseline for further research in this area.

Ground truth data collection for mobile location recognition is difficult. One approach is to mine from online photo collections like Flickr [7, 6]. Database construction from such sources is a challenging undertaking in its own right, since those images tend to be poorly labeled, disorganized, and unevenly distributed in the real world.

Another approach is to harness street-level data collected using surveying vehicles. In this case, the data are acquired by vehicle-mounted cameras with wide-angle lenses capturing spherical panoramic images. This approach offers uniform coverage of most city locations and produces imagery precisely calibrated using an inertial measurement unit. However, such datasets are rarely available outside the companies that collect them.

As part of this publication, we make a large sample of organized street-level data available to the community to help stimulate further research. The dataset is generated from 150k panoramic images of San Francisco that were aligned to a 3D model of the city consisting of 14k buildings obtained from footprint and elevation data [18]. The images were labeled by projecting the 3D model into the panoramic images, computing visibility, and recording the identities of

[1] http://www.google.com/mobile/goggles

visible buildings. Precise camera calibration data is provided for all the database images. We also release several hundred query images that were captured a few months later using a variety of mobile phones and labeled with the same building IDs as the database images. The query images also have GPS information included.

Mobile location recognition is challenging for many reasons. The query images are usually taken under very different conditions than the database images. Buildings tend to have few discriminative visual features and many repetitive structures, and their 3D geometries are not easily captured by simple affine or projective transformations. Finally, outdoor urban environments are very dynamic and undergo many changes in appearance over time.

We present an image retrieval pipeline that is a result of a comprehensive study of the performance of the system using a variety of algorithms and techniques. In particular, we improve the retrieval performance relative to a baseline vocabulary tree [15] using histogram equalization and upright feature keypoints. Furthermore, we develop an effective method of incorporating priors on the user's position (e.g. using network cell or GPS coordinates associated with a query image) and show that the same data structure can be used for both GPS-aware and GPS-agnostic schemes.

All these contributions are integrated and evaluated on viewpoint-aligned (e.g. [26, 20]) and facade-aligned (e.g. [19, 2]) database representations. We show that the two representations contain complementary information that, when properly fused, can improve the recognition results significantly. In general, we obtain a retrieval performance that is competitive with previously presented results but at a much larger scale. We feel that the proposed solution could be used as the benchmark for further studies in this area.

2. Related Work

Visual landmark identification is closely related to the image retrieval (e.g. [9, 16]), object recognition (e.g. [21, 15]) or location recognition (e.g. [25, 24, 19, 26, 20, 14, 2, 11]) problems. A commonly adopted scheme extracts local image features (e.g. [13]), quantizes their descriptors to visual words, and applies methods from text search for image retrieval [21, 15].

The closest works to ours are probably by Schindler et al. [20] and Baatz et al. [2]. Both approaches adopt vocabulary trees for image-based localization in moderately-sized image databases of a few ten thousand images. Since other city-scale location recognition data sets are not publicly available, it is difficult to compare the absolute numbers. However, we believe that our approach is competitive as it outperforms the baseline reference implementations presented in [20] and [2]. Schindler et al. work on unmodified perspective images, while Baatz et al. rectify the query images and reduce the problem to orthophoto matching on facades. One contribution of our work is that we combine both ideas and show that they contain complementary information which, when properly fused, boosts the overall retrieval rate. On top of that, and to allow future methods to evaluate and compare their approaches to ours, we publish both the database images and query images used in our experiments. To our best knowledge, no such systematically captured city-scale data sets with ground truth landmark visibility information are publicly available yet.

In our work, the visual database consists of omnidirectional images while the query images are captured using traditional perspective cameras. In prior work, typically both image groups are captured using the same camera model: for example, perspective cameras are used in [20], and omnidirectional cameras are used in [25, 14]. In our experience, directly matching perspective query images to a database of omnidirectional panoramas leads to poor performance. Instead, we use an approximate 3D model of the city [18] to convert the omnidirectional panoramas to a set of perspective images of visible buildings. This has an additional benefit of eliminating most database features that are not related to any buildings which we wish to recognize.

Torralba et al. [24] represent places using low-dimensional global image representation and use a Hidden Markov Model to solve the localization problem. However, global descriptors are not very robust to occlusions, clutter, and changes in viewpoint or orientation. Cummins and Newman [4] propose a probabilistic approach to the problem of recognizing places based on co-occurences of visual words and learn a generative model of place appearance. Also, recent work on 3D reconstruction from a large photo collection is related. Agarwal et al. [1] present a system for distributed matching and 3D reconstruction of city-scale datasets consisting of 150k images. Frahm et al. [5] leverage geometric and appearance constraints to obtain a highly parallel implementation for dense 3D reconstruction from unregistered massive photo collections. Li et al. [11] leverage feature correspondences obtained from 3D reconstructions estimated from large Internet photo collections to rank feature importance and to quickly find database features matching the query features.

For GPS-assisted landmark recognition, Takacs et al. [22] use the GPS signal to retrieve only images falling in nearby location cells. In their system, a different k-d tree is constructed for each cell. Similarly, Kumar et al. [10] use GPS to search local vocabulary trees for the purpose of visual loop closing. These prior works have been evaluated on databases containing thousands of images, whereas we study the problem of GPS-aware image retrieval from a city-scale database with millions of images. Also, we design our system so that the same data structures can be used whether or not GPS information is available.

3. Database Construction

Data is collected using a mobile mapping vehicle composed of 360° LIDAR sensor (Velodyne HDL-64E), panoramic camera (Ladybug 3), high-definition cameras (Prosilica), Global Positioning System (GPS), Inertial Measurement Unit (IMU) and Distance Measurement Instrument (DMI). The Velodyne LIDAR sensor consists of 64 lasers mounted on upper and lower blocks of 32 lasers each and the entire unit spins. This design allows for 64 separate lasers to each fire thousands of times per second, generating over one million points per second. The Ladybug 3 covers more than 80 percent of a full sphere, with six high quality 1600x1200 Sony CCD sensors capturing 8MP panoramas at 15 frames per second. All of these sensors are georeferenced through GPS and IMU.

We are working with a dataset collected for San Francisco that contains approximately 150k panoramic images captured at 4-meter intervals that are then converted to approximately 1.7 million perspective images. The building outlines come from Sanborn Inc. and consist of 2D polygons that represent the building footprint as seen from an aerial view. For each 2D polygon, base and roof elevation are also provided. If a building has complex geometry, it is very coarsely approximated as a set of sub-buildings, each having separate elevation data.

We correct the misalignment between the panoramas and the building outlines by aligning the LIDAR point clouds to the building outlines and then applying the same transformation to the camera poses using the method proposed by Pylvänäinen et al. [18]. Once we have aligned the LIDAR data and the panoramas to the building outlines, we generate the *visibility masks*. A unique color ID is assigned to each building outline, as shown in Fig. 2(b). For each panorama, the visibility mask is generated by rendering a spherical projection of building outlines using their unique color identifiers. Only the buildings that cover at least 2% of the panorama are selected as visible.

3.1. From Panoramas to Perspective Images

Matching query images directly to panoramas yields poor results because the spherical projection alters the locations and the descriptors of local image features. To address this problem, we convert the portions of panoramas containing visible buildings into perspective images. This process is illustrated in Fig. 2. For each visible building, we compute a bounding box of its projection in the visibility mask and generate overlapping perspective views of its inside. Each image has a 60° field of view, 640×480 resolution and 50% overlap with the neighboring images. Because the images are generated from the center of the panorama, we call them *perspective central images*, or PCIs. In total, we generate 1.06M PCIs.

For each PCI, we also generate a *perspective frontal image*, or a PFI, by shooting a ray through the center of projection of a PCI and computing the ray intersection point with the scene geometry. The coordinates of the intersection point $\mathbf{p}$, its normal direction $\mathbf{n}$ and the distance d to it are used to compute the location of the "frontal view." The frontal view will be generated by looking at a plane along its normal direction $\mathbf{n}$ from a location $\mathbf{f} = \mathbf{p} + d\mathbf{n}$ that is distance d away. A PFI will be generated only if an intersection point is found and the angle between the viewing direction and the normal at the intersection point is less than 45°. In total, we generate 638k PFIs.

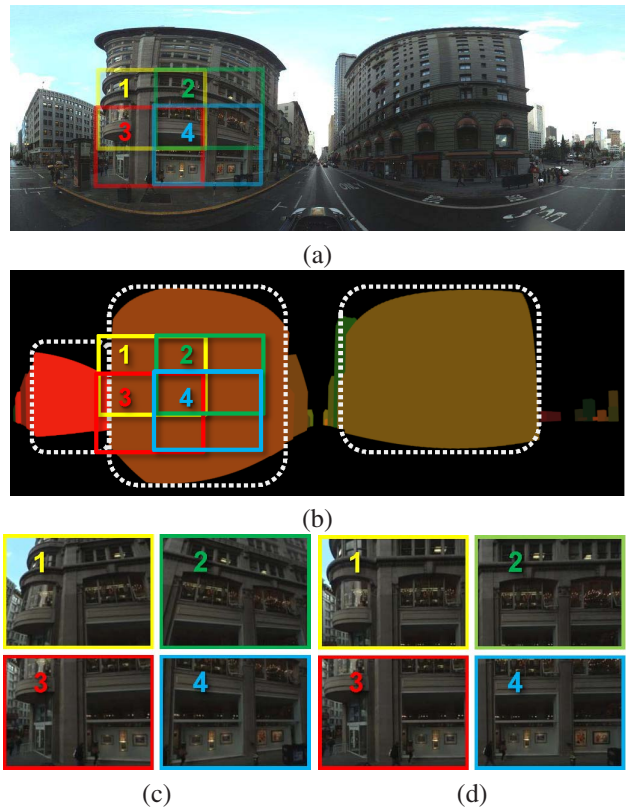


Figure 2. (a) Panoramic images are captured using a surveying vehicle. (b) The corresponding building mask of the panorama images are generated from 3D building models of the city. Visible buildings are selected if the area of the building mask exceeds a certain size, as shown by the white boxes. Views of a the visible buildings are sampled at a regular interval with an overlap of 50% (example views are shown in numbered color boxes in (a)). Two different images are then generated from each view: (c) the perspective central image, and (d) perspective frontal image.

3.2. Histogram Equalization

In urban environments, buildings can cast shadows on each other. This is the case exhibited in Fig. 1(a), where the left half of the street shows much greater brightness and contrast in the panorama. Most interest point detectors, including the difference-of-Gaussian (DoG) interest point

detector that we use[2], are not robust against large changes in contrast and brightness, or at least the standard parameters do not work well and manual tweaking with respect to the image content would be required to obtain a reasonable number of detections. In [12], the authors suggest to use a locally adaptive threshold for feature keypoints, but we prefer a computationally simpler solution for processing a large database of images.

We have found histogram equalization [8] to be effective in adaptively enhancing the contrast of all the images prior to feature extraction. As we will show in Sec. 5, histogram equalization as a pre-processing step before feature extraction noticeably improves retrieval accuracy.

3.3. Upright Feature Keypoints

The rotational invariance in local image features comes at a loss in distinctiveness. As shown in [2, 3], upright feature keypoints can be more discriminative than oriented keypoints, provided that the two images being compared have roughly the same global orientation. In constructing the database, we can constrain all the images to be approximately gravity-aligned. For each query photo, we can exploit the fact that the vast majority of users take photos in portrait and landscape mode. Using the mobile device's motion sensor readings, we rotate each query photo by a multiple of 90° so that the photo becomes approximately gravity-aligned (for some of the most recent and future devices, gravity direction is available even with higher accuracy than this 90° quantization, making the upright representation even more attractive). As shown in Sec. 5, the greater discriminative capability of upright keypoints leads to improved retrieval. Upright keypoints are also faster to compute than oriented keypoints, enabling real-time augmented reality applications as in [23].

3.4. Query Images

We test retrieval performance using a set of 803 query images of landmarks in San Francisco taken with several different camera phones by various people several months after the database images were collected. These images are taken from a pedestrian's perspective at street level. A few examples from this query set are shown in Fig. 3. Challenges present in these queries include clutter (e.g. vehicles, pedestrians, seasonal vegetation), shadows cast on buildings, reflections and glare on windows, and severe perspective with extreme angles (e.g. photos taken at street corners). There are often large photometric and geometric distortions separating these query images from their closest matches in the database. For 596 of the query images, real GPS coordinates are collected from the camera phones' sensors. The remaining 207 query images have simulated GPS coordinates, as described in Sec. 4.1.

[2] http://www.cs.unc.edu/~ccwu/siftgpu

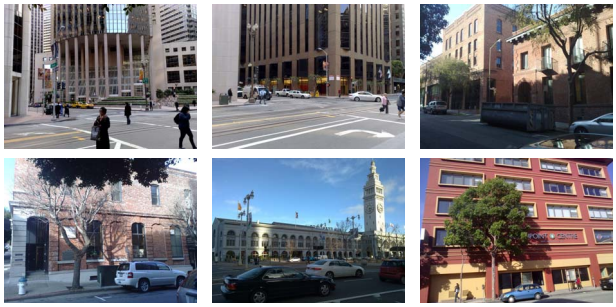


Figure 3. Examples from the set of 803 query images taken with a variety of different camera phones by several different people in San Francisco.

3.5. Public Availability

The entire database is made publicly available online.[3] The database contains 1.7 million perspective images (PCIs and PFIs) generated by the process described in Sec. 3.1. For each PCI, we provide the field of view, center of projection, camera orientation, visibility mask, and building label. For each PFI, we additionally give the warping plane parameters: $\mathbf{p}$, $\mathbf{n}$ and d. We also make 803 cell phone query images available as part of the database. For each query image, we provide its building label and GPS tag, and we specify if the GPS tag is real or simulated.

4. Recognition Pipeline

An overview of our recognition pipeline is shown in Fig. 4. When a query image is taken, it is processed in two parallel pipelines, one for PCIs and the other for PFIs. In the PCI pipeline, the database PCIs are scored using a vocabulary tree trained on SIFT descriptors, geographically distant landmarks are excluded using GPS coordinates associated with the query image (when GPS is available), and geometric verification is performed on a shortlist of database candidates. The PFI pipeline works analogously. Before matching against PFIs, however, the query image must be rectified by detecting line segments[4] and then estimating vanishing points [2]. The other important difference between the PCI and PFI pipelines is in the geometric verification stage: PCIs uses RANSAC with a 2D affine model, while PFIs use a 3 degree-of-freedom (DOF) scale and offset check. Finally, the results of both pipelines are merged, taking advantage of the complementary information of PCIs and PFIs. This yields a hybrid system that performs noticeably better than either the PCI or PFI system can alone.

4.1. GPS-Constrained Database Search

Many mobile devices now embed GPS receivers. Although the GPS readings can be inexact in urban environ-

[3] http://www.nn4d.com/sanfranciscolandmark

[4] http://www.cs.illinois.edu/homes/dhoiem/software

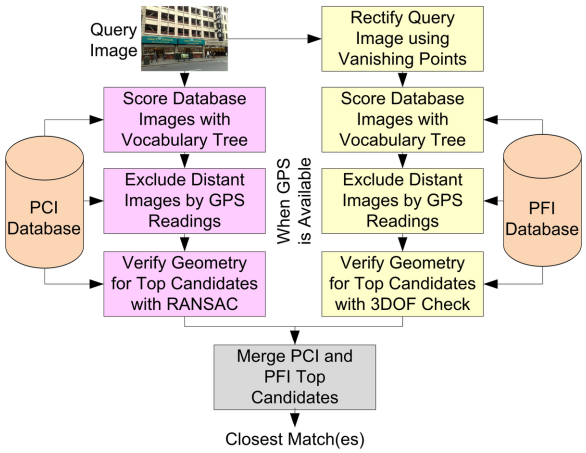
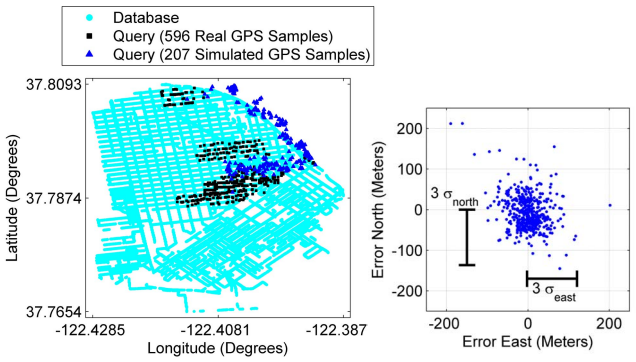


Figure 4. Pipeline for our city-scale location recognition system.



(a) (b)

Figure 5. (a) Locations of 1.7 million database images and 803 query images. All 803 query images are taken with actual camera phones in downtown San Francisco. 596 of the query images had real GPS readings associated with them. GPS readings were simulated for the remaining 207 query images using a Gaussian error model estimated from the real GPS readings. (b) Errors in 596 real GPS coordinates.

ments, even a coarse GPS estimate can be very helpful in narrowing the search space in a large image database. The locations of our database images are plotted in Fig. 5(a). Also shown in Fig. 5(a) are the locations where 803 query images were taken with various camera phones. Of the 803 query images, 596 of the images (74 percent) have real GPS readings captured by the onboard GPS sensors. Errors in these GPS readings relative to the ground truth locations (computed as the centers of the ground truth buildings) are plotted in Fig. 5(b), where it can be observed that most errors are confined to ± 150 meters. To facilitate experimentation on all 803 query images, we created a two-dimensional Gaussian model from the GPS error samples and simulated GPS readings for 207 query images which did not have actual GPS recordings. The query images with actual and simulated GPS readings are separately labeled in Fig. 5(a).

GPS-constrained landmark image retrieval can be performed in two ways. Either the database is divided into different location cells and a separate approximate nearest neighbor (ANN) tree is trained for each cell [22], or a single ANN tree is trained for the whole city and GPS simply determines which database images are scored. The disadvantages of the first method are that:

- In a practical system, two separate databases must be maintained on the server, for queries sent with and without GPS information, thereby doubling the memory requirement.
- If there are N_{cells} location cells and the total memory allocated to the ANN trees is M_{trees}, the tree for each cell can only use $M_{\text{trees}}/N_{\text{cells}}$ memory on average, thereby limiting the size (e.g. depth of an ANN tree) and distinctiveness of each tree's vocabulary.
- There may be boundary issues when transitioning between location cells.

The disadvantages of the first method motivate us to pursue the alternative approach. When a GPS reading is available, one city-scale tree is searched but only the database images nearby to the current GPS position are scored. Besides avoiding the drawbacks of the first method, several other advantages are gained by this approach:

- The single tree can use the full M_{trees} memory. In experiments, we found that using a large tree for the whole city gives better retrieval performance than using many small trees for separate location cells.
- Although we use a hard cutoff for the geographic search neighborhood here, we note that database images can be included according to a probability distribution (e.g. Gaussian) centered on the current GPS coordinate.
- For improved performance, dynamic adjustment of the neighborhood size is possible when there exists reliable information about the GPS error variance.

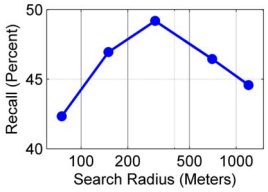


Figure 6. Recall for a shortlist of top 50 database candidates out of 1.06 million PCIs as a function of the spatial search radius.

Given noisy GPS coordinates $\mathbf{Q} = (Q_{\text{lat}}, Q_{\text{long}})$ for a query image, we score the database images that are within R meters of $\mathbf{Q}$. There is an interesting trade-off that can be observed as the value of R varies, as illustrated in Fig. 6. For small values of R, fewer incorrect database candidates are considered, but it is more likely for the correct database

candidates to be excluded when there are errors in the GPS signal **Q**. The converse is true for large values of R. We find there is an optimal value of $R \approx 300$ meters, where the best trade-off between excluding incorrect and including correct database images is obtained. Note that GPS alone is not enough for landmark identification, as there are 230 different buildings on average within a 300 meter radius.

4.2. Combining PCI and PFI Results

From the shortlists of database candidates obtained from the PCI and PFI pipelines, we wish to produce a common ranking of all the candidates. Suppose the inlier counts after geometric verification (GV) for the top n candidates are $\{N_{\text{pci},1}, \cdots, N_{\text{pci},n}\}$ for PCIs and $\{N_{\text{pfi},1}, \cdots, N_{\text{pfi},n}\}$ for PFIs. Although the GV methods are different for the PCIs and PFIs, we can make the two types of inlier counts directly comparable by multiplying the PFI inlier counts by a parameter α. Then, we sort the concatenated list $\{N_{\text{pci},1}, \cdots, N_{\text{pci},n}, \alpha N_{\text{pfi},1}, \cdots, \alpha N_{\text{pfi},n}\}$ and retain the top n candidates out of the $2n$ candidates. Because the PCIs and PFIs contain complementary visual information, sometimes one type better matches a query image than the other type. In the next section, we show that fusing the two shortlists produces a hybrid system that noticeably outperforms either pipeline alone, and we study which value of α optimizes the retrieval performance.

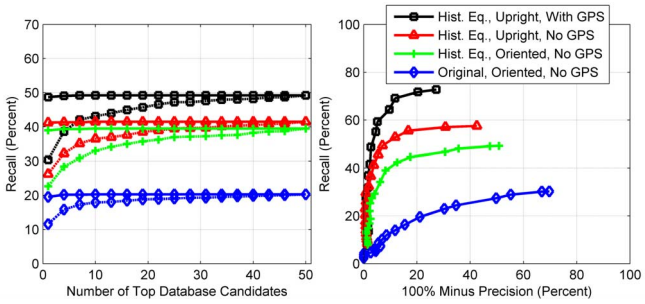
5. Experimental Results

In this section, we test the retrieval performance of our proposed recognition pipeline on the new city-scale landmark database. All retrieval schemes use DoG keypoints and SIFT descriptors. Each vocabulary tree has depth 6, branch factor 10, and 1 million leaf nodes and is trained using 16 million descriptors randomly chosen from the database. TF-IDF scoring with soft binning [17] is used.

5.1. Perspective Central Images

We measure the recall of correct buildings over the set of 803 query images as a function of the top n candidates. Fig. 7(a) plots the recall as n ranges from 1 to 50 for several different methods. In the plot, several parameters are varied to produce different combinations:

- **Hist. Eq.** refers to the PCIs undergoing histogram equalization prior to feature extraction, whereas **Original** refers to extracting features from the original PCIs.
- **Upright** refers to extracting DoG keypoints with upright orientation, whereas **Oriented** refers to extracting DoG keypoints oriented along the dominant gradient direction.
- **With GPS** refers to using a query image's GPS coordinates to consider only nearby database images, whereas **No GPS** refers to considering every image in the database.



(a) (b)

Figure 7. For database of 1.06M PCIs and 803 query images. (a) Recall versus number of top database candidates in shortlist for 95% precision. The solid and dashed curves are post and pre geometric verification (GV) results, respectively. (b) Precision versus recall for T_{PCI} ranging from 0 to 100.

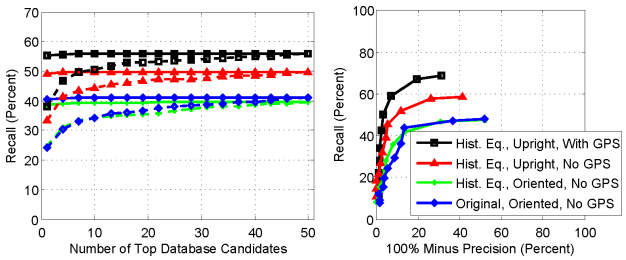
Fig. 7(a) shows the incremental advantages of different enhancements to a baseline vocabulary tree scheme. First, performing histogram equalization on the PCIs noticeably boosts the recall, because the DoG interest point detector is sensitive to the large changes in contrast that frequently occur when building facades are cast in shadows. Second, using upright keypoints helps retrieval, because the features generated from upright keypoints are more discriminative than those generated from oriented keypoints, an observation confirmed in previous work [3, 2]. Third, geographically limiting the search neighborhood using GPS coordinates associated with a query image can substantially improve recall, even on top of the two previous enhancements of histogram equalization and upright keypoints.

In our system, if the number of inliers after geometric verification is higher than a threshold T_{PCI}, we report the label of the best retrieved database image. Otherwise, we report a "no match" answer. By decreasing T_{PCI}, we can increase the recall while lowering the precision. The precision-recall tradeoff for the different schemes is shown in Fig. 7(b). A practical mobile landmark identitication system needs to operate at high precision to avoid returning false positives to users. Thus, all plots of recall-vs-top-n-candidates in this paper are generated assuming an operating point where our best scheme (*Histogram Equalized, Upright, With GPS*) achieves precision $\approx 95\%$, corresponding to $T_{PCI} = 30$ inliers for PCIs.

5.2. Perspective Frontal Images

We run experiments using the PFI pipeline and measure recall in the same way as in the preceding section. For geometric verification in the PFI pipeline, we use an inlier threshold of $T_{PFI} = 25$ to reach the target precision of 95%. We also analyze the same combinations of parameters

(see Fig. 8). Generally, using PFIs improves retrieval compared to PCIs because the geometric distortions between the query and database images have been reduced. This is most noticable for the *Original, Oriented, No GPS* case, but even for the other cases, the gain is in the range of 5–10%.



(a) (b)

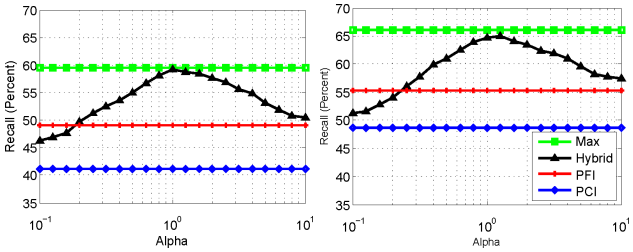
Figure 8. For database of 638k PFIs and 803 query images. (a) Recall versus number of top database candidates in shortlist for 95% precision. The solid and dashed curves are post and pre GV results, respectively. (b) Precision versus recall for T_{PFI} ranging from 0 to 100.

5.3. Combined PCI and PFI Retrieval

Since the geometric verification methods used for PCIs and PFIs are different, we pre-multiply the PFI inlier counts by a tuning parameter α. First, we determine the optimal value of α. We look at the recall of the top one image after geometric verification for α varying between 0.1 and 10 (see Fig. 9). For extreme values of α, recall approaches that of PCI and PFI, respectively. This confirms our intuition that in these cases the influence of one pipeline is negligible and the results are almost exclusively dictated by the other. Interestingly, the optimum for our best scheme (*Histogram Equalized, Upright, With GPS*) occurs approximately at $1.2 = \frac{T_{PCI}}{T_{PFI}}$, which corresponds to the ratio of the inlier thresholds for the PFI resp. PCI pipelines. We use this optimal value of α in the remaining experiments.

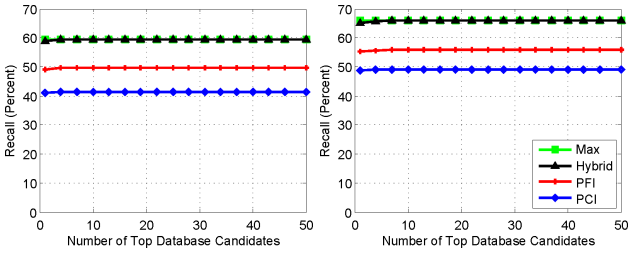
We calculate recall for a hypothetical scheme that always boosts the correct candidate from the two shortlists to the first place. In the graphs we denote this as the *Max*, since no method of combining the two shortlists can beat this. We see that our proposed scheme is in effect very close to the theoretical maximum.

We compute the same type of recall-vs-top-n-candidates curves for the hybrid scheme as in the previous sections. For reference, we also include the *Max* and the underlying PCI and PFI curves. For both plots, we use histogram equalization and upright features, once with and once without GPS (Fig. 10). The hybrid scheme noticeably boosts recall compared to either PCIs or PFIs, by about 10% for both the GPS-aware and GPS-agnostic modes.



(a) (b)

Figure 9. Recall of the top one image as a function of α. (a) Without GPS. (b) With GPS. In both cases, the optimum lies close to $\alpha = 1.2$ and gets within less than 1% of the theoretical maximum.



(a) (b)

Figure 10. Recall for the hybrid method as a function of the number of candidates in the shortlist. (a) Without GPS. (b) With GPS. The curves labeled *PCI* and *PFI* are the same as in Figs. 7(a) and 8(a) the curves labeled *Hist. Eq., Upright, With/No GPS*.

6. Conclusion

Image-based landmark identification with mobile devices remains an active area of research. Carefully organized and labeled city-scale datasets, however, are difficult to obtain in practice. One of the main contributions of this paper is making publicly available a database that provides dense and spatially well distributed coverage of San Francisco landmarks. The database contains 1.7 million images with precise ground truth labels and geotags. A set of query images taken with different camera phones is also provided to test retrieval performance. Our other major contribution is a strong benchmark on the new dataset, consisting of a set of repeatable experiments using state-of-the-art image retrieval techniques. Several methods are developed to significantly improve retrieval performance: histogram equalizing the database images to counter the effect of shadows, extracting upright feature keypoints for more distinctive description, utilizing GPS for filtering database candidates after vocabulary tree scoring, and combining perspective and view-invariant retrieval results to the get the best of both types of images. Compared to a baseline vocabulary tree, our enhancements yield over 45% improvement in recall, from 20% (Fig. 7(a), blue curve) to 65% (Fig. 10(b), black curve). We hope that the new dataset and our initial benchmark will stimulate further research in this area.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing Rome. *Computer*, 43:40–47, 2010. 738

[2] G. Baatz, K. Koeser, D. Chen, R. Grzeszczuk, and M. Pollefeys. Handling Urban Location Recognition as a 2D Homothetic Problem. In *ECCV '10: Proceedings of the 11th European Conference on Computer Vision*, 2010. 738, 740, 742

[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008. 740, 742

[4] M. Cummins and P. Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *International Journal of Robotics Research*, 27(6):647–665, 2008. 737, 738

[5] J.-M. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building Rome on a Cloudless Day. In *ECCV '10: Proceedings of the 11th European Conference on Computer Vision*, 2010. 738

[6] S. Gammeter, D. Tingdahl, T. Quack, and L. V. Gool. Size Does Matter: Improving Object Recognition and 3D Reconstruction with Cross-Media Analysis of Image Clusters. In *ECCV '10: Proceedings of the 11th European Conference on Computer Vision*, 2010. 737

[7] L. V. Gool, M. D. Breitenstein, S. Gammeter, H. Grabner, and T. Quack. Mining from Large Image Sets. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, 2009. 737

[8] R. A. Hummel. Histogram Modification Techniques. *Computer Graphics and Image Processing*, 4(3):209 – 224, 1975. 740

[9] H. Jegou, M. Douze, and C. Schmid. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 304–317, 2008. 738

[10] A. Kumar, J.-P. Tardif, R. Anati, and K. Daniilidis. Experiments on Visual Loop Closing using Vocabulary Trees. In *Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 738

[11] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition using Prioritized Feature Matching. In *ECCV '10: Proceedings of the 11th European Conference on Computer Vision*, 2010. 738

[12] A. Lingua, D. Marenchino, and F. Nex. Performance Analysis of the SIFT Operator for Automatic Feature Extraction and Matching in Photogrammetric Applications. *Sensors*, 9(5):3745–3766, 2009. 740

[13] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 738

[14] A. C. Murillo, J. J. Guerrero, and C. Sagues. SURF Features for Efficient Robot Localization with Omnidirectional Images. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3901–3907, 2007. 738

[15] D. Nister and H. Stewenius. Scalable Recognition with a Vocabulary Tree. In *Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, 2006. 738

[16] M. Perďoch, O. Chum, and J. Matas. Efficient Representation of Local Geometry for Large Scale Object Retrieval. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9–16, June 2009. 738

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 742

[18] T. Pylvänäinen, K. Roimela, R. Vedantham, J. Itaranta, R. Wang, and R. Grzeszczuk. Automatic Alignment and Multi-View Segmentation of Street View Data using 3D Shape Priors. In *Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2010. 737, 738, 739

[19] D. Robertson and R. Cipolla. An Image-Based System for Urban Navigation. In *In Proc. of British Machine Vision Conference (BMVC'04)*, pages 7–9, 2004. 738

[20] G. Schindler, M. Brown, and R. Szeliski. City-Scale Location Recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, 2007. 738

[21] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the International Conference on Computer Vision*, volume 2, pages 1470–1477, Oct. 2003. 738

[22] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.-C. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod. Outdoors Augmented Reality on Mobile Phone using Loxel-Based Visual Feature Organization. In *MIR '08: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 427–434, 2008. 738, 741

[23] G. Takacs, V. Chandrasekhar, S. Tsai, D. Chen, R. Grzeszczuk, and B. Girod. Unified Real-Time Tracking and Recognition with Rotation-Invariant Fast Features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 740

[24] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based Vision System for Place and Object Recognition. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 273, 2003. 738

[25] I. Ulrich and I. Nourbakhsh. Appearance-based Place Recognition for Topological Localization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation, ICRA'00*, pages 1023–1029, 2000. 738

[26] W. Zhang and J. Kosecka. Image Based Localization in Urban Environments. In *3DPVT '06: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*, pages 33–40, 2006. 738