24/7 place recognition by view synthesis

Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, Tomas Pajdla

24/7 place recognition by view synthesis

Akihiko Torii
Tokyo Tech*

Relja Arandjelović
INRIA†

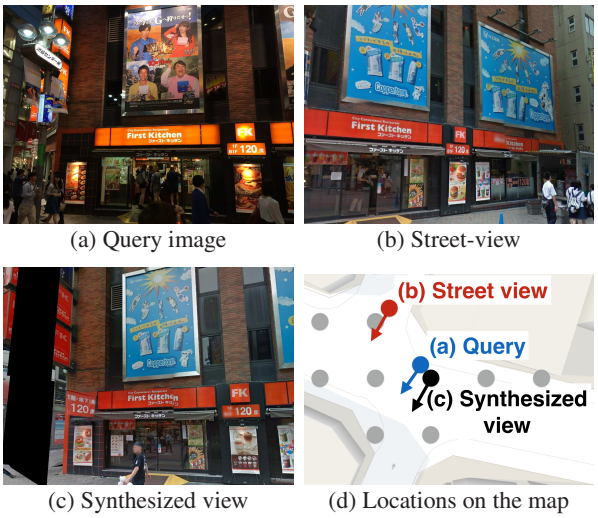
Josef Sivic
INRIA†

Masatoshi Okutomi
Tokyo Tech*

Tomas Pajdla
CTU in Prague‡

Abstract

We address the problem of large-scale visual place recognition for situations where the scene undergoes a major change in appearance, for example, due to illumination (day/night), change of seasons, aging, or structural modifications over time such as buildings built or destroyed. Such situations represent a major challenge for current large-scale place recognition methods. This work has the following three principal contributions. First, we demonstrate that matching across large changes in the scene appearance becomes much easier when both the query image and the database image depict the scene from approximately the same viewpoint. Second, based on this observation, we develop a new place recognition approach that combines (i) an efficient synthesis of novel views with (ii) a compact indexable image representation. Third, we introduce a new challenging dataset of 1,125 camera-phone query images of Tokyo that contain major changes in illumination (day, sunset, night) as well as structural changes in the scene. We demonstrate that the proposed approach significantly outperforms other large-scale place recognition techniques on this challenging data.



(a) Query image (b) Street-view (c) Synthesized view (d) Locations on the map

Figure 1. **Matching across major changes in scene appearance is easier for similar viewpoints.** (a) Query image. (b) The original database image cannot be matched to the query due to a major change in scene appearance combined with the change in the viewpoint. (c) Matching a more similar synthesized view is possible. (d) Illustration of locations of (a-c) on the map. The dots and arrows indicate the camera positions and view directions.

1. Introduction

Recent years have seen a tremendous progress [3, 6, 7, 8, 10, 14, 24, 28, 34, 35, 36, 40, 44] in the large-scale visual place recognition problem [27, 36]. It is now possible to obtain an accurate camera position of a query photograph within an entire city represented by a dataset of 1M images [3, 8, 40] or a reconstructed 3D point cloud [28, 34]. These representations are built on local invariant features such as SIFT [29] so that recognition can proceed across moderate changes in viewpoint, scale or partial occlusion by other objects. Efficiency is achieved by employing inverted file [33, 39] or product quantization [20] indexing techniques. Despite this progress, identifying the same place across major changes in the scene appearance due to illumination (day/night), change of seasons, aging, or structural modifications over time [12, 30], as shown in figure 1, remains a major challenge. Solving this problem would have, however, significant practical implications. Imagine, for example, automatically searching public archives to find all imagery depicting the same place to analyze changes over time for applications in architecture, archeology and urban planning; or visualize the same place in different illuminations, seasons or backward in time.

In this paper, we demonstrate that matching across large changes in scene appearance is easier when both the query image and the database image depict the scene from approximately the same viewpoint. We implement this idea by synthesizing virtual views on a densely sampled grid on the map. This poses the following three major challenges.

*Department of Mechanical and Control Engineering, Graduate School of Science and Engineering, Tokyo Institute of Technology

†WILLOW project, Laboratoire d'Informatique de l'École Normale Supérieure, ENS/INRIA/CNRS UMR 8548.

‡Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Enginnering, Czech Technical University in Prague

First, how can we efficiently synthesize virtual viewpoints for an entire city? Second, how do we deal with the increased database size augmented by the additional synthesized views? Finally, how do we represent the synthetic views in a way that is robust to the large changes in scene appearance?

To address these issues, we, first, develop a view synthesis method that can render virtual views directly from Google street-view panoramas and their associated approximate depth maps, not requiring to reconstruct an accurate 3D model of the scene. While the resulting images are often noisy and contain artifacts, we show that this representation is sufficient for the large-scale place recognition task. The key advantage of this approach is that the street-view data is available world-wide opening-up the possibility for a truly planet-scale [23] place recognition. Secondly, to cope with the large amount of synthesized data – as much as nine times more images than in the original street-view – we use the compact VLAD encoding [2, 21] of local image descriptors, which is amenable to efficient compression, storage and indexing. Finally, we represent images using densely sampled local gradient based descriptors (SIFT [29] in our case) across multiple scales. We found that this representation is more robust to large changes in appearance due to illumination, aging, *etc*. as it does not rely on repeatable detection of local invariant features, such as the Laplacian of Gaussian [29]. While local invariant features have been successfully used for almost two decades to concisely represent images for matching across viewpoint and scale [41] they are often non-repeatable across non-modeled changes in appearance due to, *e.g*. strong perspective effects or major changes in the scene illumination [4, 9]. Not relying on the local invariant keypoint detection comes at a price of reduced invariance to geometric transformation. However, we have found this is in fact an advantage, rather than a problem, as the resulting representation is more distinctive and thus copes better with the increased rate of false positive images due to the much larger database augmented with synthetic views.

2. Related work

Place recognition with local-invariant features. The large-scale place recognition is often formulated as a variation of image retrieval [22, 33] where the query photograph is localized by matching it to a large database of geo-tagged images such as Google street-view [6, 8, 10, 14, 24, 35, 36, 40, 44]. The 3D structure of the environment can be also reconstructed beforehand and the query is then matched directly to the reconstructed point-cloud [28, 34] rather than individual images. The underlying appearance representation for these methods is based on local invariant features [41], either aggregated into an image-level indexable representation [8, 10, 14, 24, 40, 44], or associated to individual reconstructed 3D points [28, 34]. These methods have shown excellent performance for large-scale matching across moderate changes of scale and viewpoint that are modeled by the local invariant feature detectors. However, matching across non-modeled appearance variations such as major changes in illumination, aging, or season are still a challenge.

We investigate compact representations based on descriptors densely sampled across the image rather than based on local-invariant features. Densely sampled descriptors have been long used for category-level recognition [5, 11, 26, 32] including category-level localization [14], but due to their limited invariance to geometric transformations have been introduced to instance-level recognition only recently [45]. While we build on this work, we show that combining dense representations with virtual view synthesis can be used for large-scale place recognition across significant changes of scene appearance.

Virtual views for instance-level matching. Related to our work are also methods that generate some form of virtual data for instance-level matching, but typically they focus on extending the range of recognizable viewpoints [17, 37, 43] or matching across domains [4, 38] and do not consider compact representations for large-scale applications. Irschara *et al.* [17] generate bag-of-visual-word descriptors extracted from existing views for virtual locations on a map to better model scene visibility. Shan *et al.* [37] use 3D structure to synthesize virtual views to match across extreme viewpoint changes for alignment of aerial to ground-level imagery. Wu *et al.* [43] locally rectify images based on the underlying 3D structure to extend the viewpoint invariance of local invariant features (SIFT). Their method has been successfully applied for place recognition [8] but requires either known 3D structure or rectification on the query side. Recently, rendering virtual views has been also explored for cross-domain matching to align paintings to 3D models [4] or to match SIFT descriptors between images and laser-scans [38].

Modelling scene illumination for place recognition. In place recognition, the related work on modeling outdoor illumination has focused on estimating locations and time-stamps from observed illumination effects [13, 18]. In contrast, we focus on recognizing the same scene across changes of illumination. However, if illumination effects could be reliably synthesized [25] the resulting imagery could be used to further expand the image database.

3. Matching local descriptors across large changes in appearance

In this section we investigate the challenges of using local invariant features for image matching across major changes

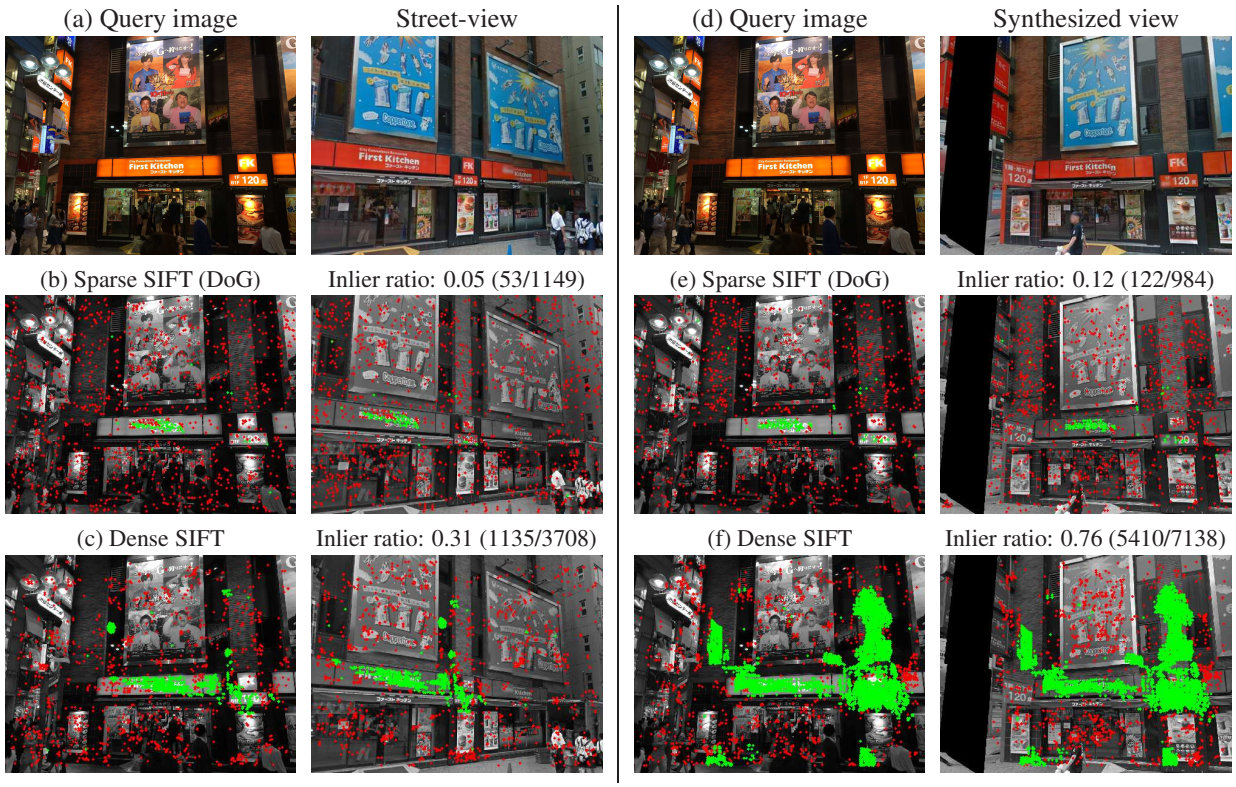


Figure 2. **Matching across illumination and structural changes in the scene. First row:** The same query image is matched to a street-view image depicting the same place from a different viewpoint (a) and to a synthesized virtual view depicting the query place from the same viewpoint (d). **Second row:** Matching sparsely sampled SIFT descriptors across a major change in illumination is difficult for the same (e) as well as for the different (b) viewpoints. **Third row:** Densely sampled descriptors can be matched across a large change in illumination (c) and the matching is much easier when the viewpoint is similar (f). In all cases the tentative matches are shown in red and geometrically verified matches are shown in green. Note how the proposed method (f), based on densely sampled descriptors coupled with virtual view synthesis, obtains significantly higher inlier ratio (0.76) on this challenging image pair with major illumination and structural changes in the scene.

in scene appearance due to day/night illumination and structural changes in the scene. We first illustrate that local invariant features based on the difference of Gaussian (DoG) feature detector are not reliably repeatable in such conditions. Then we show that densely sampled descriptors result in better matches, but suffer from limited invariance to geometric transformations (scale and viewpoint). Finally, we demonstrate that matching can be significantly improved when we match to a virtual view synthesized from approximately the same viewpoint. In this section we illustrate the above points on a matching example shown in figure 2. We verify these findings quantitatively on the place recognition task in section 5.

In all examples in figure 2 we build tentative matches by finding mutually nearest descriptors. The tentative matches are shown in red. We then geometrically verify the matches by repeatedly finding several homographies using RANSAC. The geometrically consistent matches (inliers) are shown in green. We deem all geometrically verified matches as correct (though few incorrect matches may remain). The quality of matching is measured by the inlier ratio, *i.e*. the proportion of geometrically consistent matches. The inlier ratio is between 0 and 1 with a perfect score of 1 when all tentative matches are geometrically consistent.

First, we match the upright RootSIFT descriptors [1] sampled at DoG keypoints [29] between a query image and a street-view image depicting the query place (figure 2(a)) from a different viewpoint. The matches are shown in figure 2(b) and result in an inlier ratio of only 0.05, clearly demonstrating the difficulty of matching DoG keypoints across large changes in appearance.

Second, we repeat the same procedure for the synthesized view (figure 2(d)), which captures the query place from approximately the same viewpoint as the query image. The result is shown in figure 2(e). The resulting inlier ratio of only 0.12 indicates that matching the DoG keypoints across large changes in appearance is difficult despite the fact that the two views have the same viewpoint.

Third, we extract RootSIFT descriptors with a width of 40 pixels (in a 640×480 image) on a regular densely sam-

pled grid with a stride of 2 pixels. The descriptor matching was performed in the same manner as for the descriptors extracted at the sparsely detected keypoints. Matching the densely sampled descriptors across different viewpoints and illuminations already shows an improvement compared to sparse keypoints, with the inlier ratio increasing from 0.05 to 0.31 (figure 2(c)). The fact that the descriptor (SIFT) is identical for both sampling methods suggests that the main problem is non-repeatability of the Difference of Gaussian local invariant features underpinning the sparsely sampled method, rather than the descriptor itself.

Finally, we apply the densely sampled descriptors to the image pair with different illuminations but similar viewpoints (figure 2(d)). The matches are shown in figure 2(f). The inlier ratio further increases to 0.76 clearly demonstrating the benefits of virtual view synthesis for dense descriptor matching.

4. View synthesis from street-level imagery

In this section we describe our view synthesis method that expands the database of the geo-tagged images with additional viewpoints sampled on a regular grid. To synthesize additional views we use the existing panoramic imagery together with an approximate piece-wise planar depth map associated with each panorama, as illustrated in figure 4. The piece-wise planar depth map provides only a very coarse 3D structure of the scene, which often leads to visible artifacts in the synthesized imagery. However, in section 5 we demonstrate that this quality is sufficient to significantly improve place recognition performance. In addition, this data is essentially available world-wide [15], thus opening up the possibility of planet-scale view synthesis and place recognition [23]. The view synthesis proceeds in two steps. We synthesize the candidate virtual camera locations, which is followed by synthesizing individual views. The two steps are discussed next.

We generate candidate camera positions on a regular $5m \times 5m$ grid on the map that covers the original street-view camera positions. We only generate camera positions that are within $20m$ distance from the original street-view trajectory, where the trajectory is obtained by connecting the neighboring street-view camera positions. We found that going farther than $20m$ often produces significant artifacts in the synthesized views. We also use the available depth maps to discard camera positions that would lie inside buildings. The camera positions of the synthesized views are illustrated on the map in figure 3.

To synthesize the virtual views at the particular virtual camera position we use the panorama and depth map downloaded from Google maps [15]. Each panorama captures $360°$ by $180°$ horizontal and vertical viewing angle, respectively, and has the size $13,312 \times 6,656$ pixels, as illustrated in figure 4(a). The depth map is encoded as a set of 3D plane

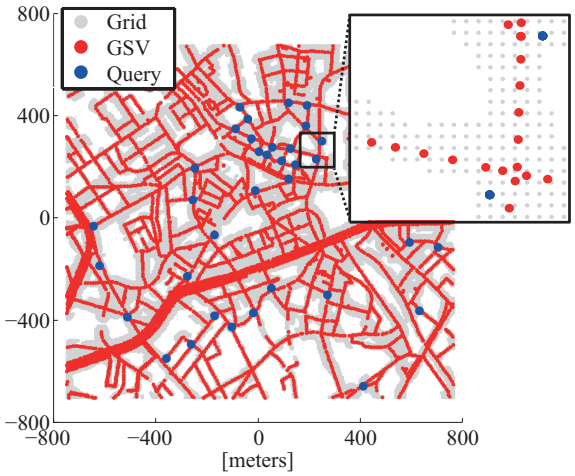
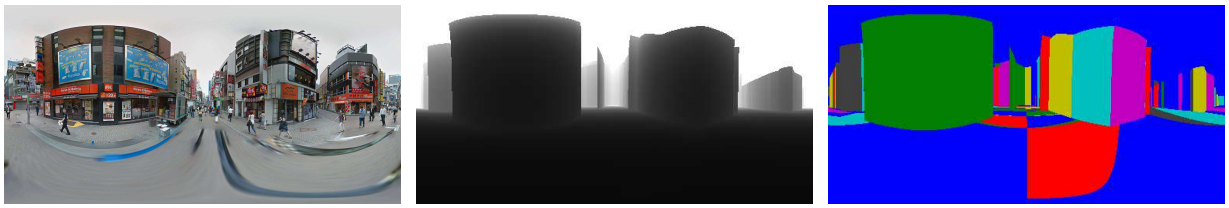


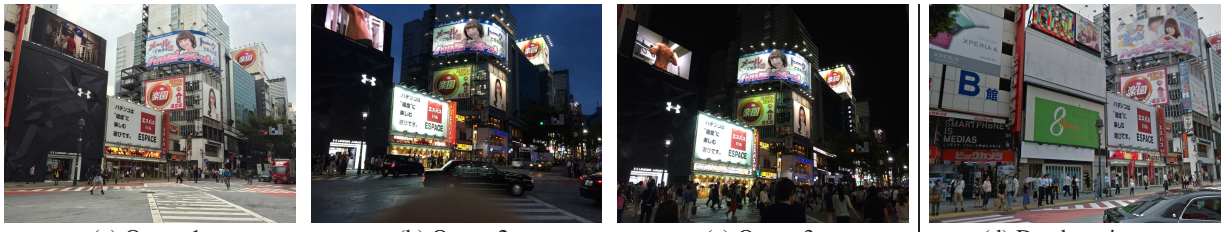
Figure 3. **Combining street-view imagery with synthetic views.** The figure shows camera positions for part of the 24/7-Tokyo dataset. The positions of the original street-view images are shown in red, the positions of synthesized views (5×5m grid) are shown in grey, and the positions of query images are shown in blue. The inset (top right) shows a close-up of one road intersection. The database of geo-tagged images includes $75,984$ views generated from the original $6,332$ street-view panoramas and $597,744$ synthesized views generated at $49,812$ virtual camera positions.

parameters (normal and distance for each plane) and an 512×256 image of indices pointing, for each pixel, to one of the planes, as illustrated in figure 4(c). Using this index we can look-up the corresponding plane for each pixel, which allows us to generate the actual depth map for the panorama, as illustrated in figure 4(b). All views at a particular virtual camera position are synthesized from the panorama and depth map of the closest street-view image. Virtual views are synthesized by standard ray tracing with bilinear interpolation. In detail, for every pixel in the synthesized virtual view, we cast a ray from the center of the virtual camera, intersect it with the planar 3D structure obtained from the depth map of the closest street-view panorama, project the intersection to the street-view panorama, and interpolate the output pixel value from the neighboring pixels. For each virtual camera location we generate 12 perspective images of $1,280 \times 960$ pixels (corresponding to 60 degrees of horizontal field of view) with a pitch direction $12°$ and the following 12 yaw directions $[0°, 30°, ..., 360°]$. This perspective view sampling is similar to e.g. [8, 40]. Examples of the synthesized virtual views are shown in figures 1, 8 and 9. While the synthesized views have missing information and artifacts (e.g. incorrectly rendered people or objects), we found this simple rendering is already sufficient to improve place recognition performance. Higher quality synthesis could be potentially obtained by combining information from multiple panoramas. Rendering one virtual view takes about a second, but we expect 1-2 or-



(a) Street-view panorama (b) Associated depth-map (c) Individual scene planes

Figure 4. **Input data for view synthesis**. (a) The street-view panorama. (b) The associated piece-wise planar depth-map. Brightness indicates distance. (c) The individual scene planes are shown in different colors.



(a) Query 1. (b) Query 2. (c) Query 3. (d) Database image

Figure 5. **Example query images from the newly collected** $24/7$ **Tokyo dataset.** Each place in the query set is captured at different times of day: (a) daytime, (b) sunset, and (c) night. For comparison, the database street-view image at a close-by position is shown in (d). Note the major changes in appearance (illumination changes in the scene) between the database image (d) and the query images (a,b,c).

ders of magnitude speed-up using a graphics processing unit (GPU). We generate the same set of perspective views for original street-view images and combine the real and virtual views into a single place recognition database. Note that virtual views are only needed for extracting the compact dense VLAD descriptors as described in section 3 and can be discarded afterwards.

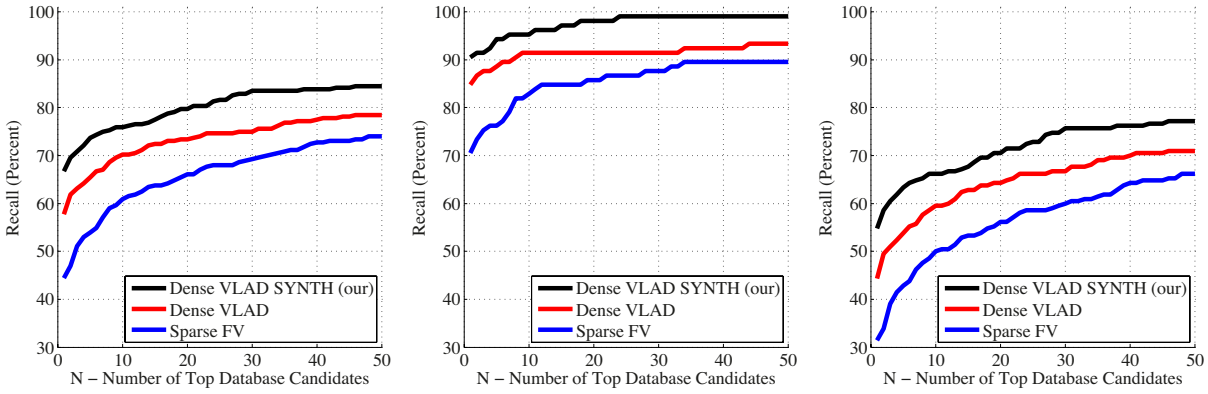
5. Experiments

In this section we describe the newly collected 24/7 Tokyo dataset, give the place recognition performance measures and outline the quantitative and qualitative results of our method compared to several baselines.

24/7 Tokyo dataset. We have collected a new test set of $1,125$ query images captured by Apple-iPhone5s and Sony-Xperia smartphones. We captured images at 125 distinct locations. At each location we captured images at 3 different viewing directions and at 3 different times of day, as illustrated in figure 5. The ground truth GPS coordinates at each location were recorded by manually localizing the position of the observer on the map at the finest zoom level. We estimate that the error of the ground truth location is below $5m$. The dataset is available at [16]. In the following evaluation, we use a subset of 315 query images within the area of about $1,600m \times 1,600m$ covered by our geo-tagged database.

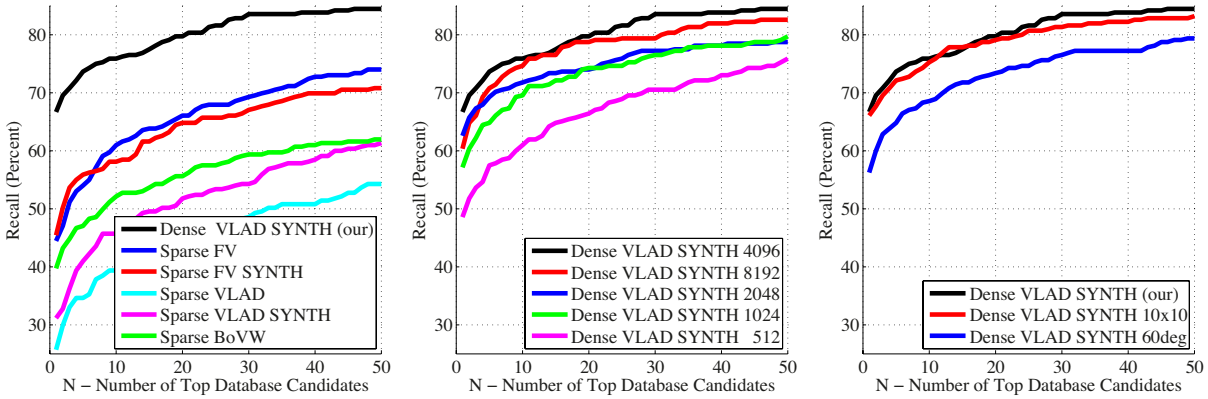
Evaluation metric. The query place is deemed correctly recognized if at least one of the top N retrieved database images is within $d = 25$ meters from the ground truth position of the query. This is a common place recognition metric used e.g. in [8, 35, 40]. The percentage of correctly recognized queries (Recall) is then plotted for different values of N.

Implementation details. To compute the Dense VLAD descriptor, we resize each image to have the maximum dimension of 640 pixels. This is beneficial for computational efficiency and limiting the smallest scale of the extracted descriptors. We extract SIFT [29] descriptors at 4 scales corresponding to region widths of 16, 24, 32 and 40 pixels. The descriptors are extracted on a regular densely sampled grid with a stride of 2 pixels. When using synthesized images, we remove descriptors that overlap with image regions that have no image data (shown in black in the synthesized imagery). We use the SIFT implementation available in Vlfeat [42] followed by the RootSIFT normalization [1], *i.e*. L1 normalization followed by element-wise square root. The visual vocabulary of 128 visual words (centroids) is built from 25M descriptors randomly sampled from the database images using k-means clustering. We have kept the original dimension of the SIFT descriptor, unlike [22]. Each image is then described by an aggregated intra-normalized [2] VLAD descriptor followed by a PCA compression to 4,096 dimensions, whitening and L2 normalization [19]. The similarity between a test query and the database images is measured by the normalized dot product, which could be efficiently performed using [20, 31]. Following [6], we diversify the returned shortlists by performing spatial non-max suppression on the map, where we associate the score from each virtual view to the closest street-view panorama.



(a) All queries (b) Day time queries (c) Sunset and night queries

Figure 6. **Evaluation on the 24/7-Tokyo dataset**. The fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis) for the proposed method (Dense VLAD SYNTH) compared to the baseline methods (Dense VLAD, Sparse FV). The performance is evaluated for all test query images (a), as well as separately for daytime queries (b), and sunset/night queries (c). The benefits of the proposed method (Dense VLAD SYNTH) is most prominent for difficult illuminations (c).



(a) Comparison with baselines. (b) Descriptor dimension (c) View sampling density

Figure 7. **Place recognition performance on the** $24/7$**-Tokyo dataset.** Each plot shows the fraction of correctly recognized queries (Recall, y-axis) vs. the number of top N retrieved database images (x-axis).

Baseline methods. We compare results to the following baselines. First, we evaluate the VLAD descriptor based on the Difference of Gaussian (DoG) local invariant features [29, 42] (Sparse VLAD). Here we use the upright RootSIFT descriptors sampled at DoG keypoints, otherwise the descriptor is constructed in the same manner as our densely sampled VLAD. Second, we compare with the standard sparse Fisher vector [22] (Sparse FV), which was shown to perform well for place recognition [40]. The Fisher vector is constructed using the same upright RootSIFT descriptors as the Sparse VLAD baseline. Following [22], the extracted SIFT descriptors are reduced to 64 dimensions by PCA. A 256-component Gaussian mixture model is then trained from 25M descriptors randomly sampled from the database images. As in [22], the resulting 256×64 dimensional Fisher vector is reduced to $4,096$ dimensions using PCA, followed by whitening and L2 normalization [19]. Finally, we also compare results to the bag-of-visual-words baseline. We construct the bag-of-visual-words descriptor (Sparse BoVW) using the same upright RootSIFT descriptors as used in the Sparse VLAD baseline. A vocabulary of 200,000 visual words is built by approximate k-means clustering [31, 33]. The resulting bag-of-visual-word vectors are re-weighted using adaptive assignment [40].

Benefits of the dense descriptor and synthesized views. First, in figure 6 we evaluate the benefits of having (i) dense descriptors (Dense VLAD) and (ii) additional synthesized views (Dense VLAD SYNTH). We compare performance with the standard Fisher vector descriptor based on local invariant features (Sparse FV), which was found to work well for place recognition [40]. We show results for all queries (figure 6(a)), but to clearly illustrate the differences we also separate the query images to daytime (figure 6(b)), and sunset/night queries (figure 6(c)). While having the

Figure 8. **Example place recognition results for our method (Dense VLAD SYNTH) compared to baseline using only sparsely sampled feature points (Sparse FV).** (Left) Query image. (2nd column) The best matching synthesized view by our method (correct). (3rd column) The best matching street-view image by the baseline (Sparse Fisher vectors without synthesized views). (4th column) The original street-view image at the closest position to the query. Note that our method can match difficult queries with challenging illumination conditions.

dense descriptor (Dense VLAD) already improves performance compared to the baseline (Sparse FV), it is the combination of the dense descriptor with synthetic virtual views (Dense VLAD SYNTH) which brings significant improvements for queries with difficult illuminations (figure 6(c)), clearly illustrating the importance of both components of our approach.

Comparison to sparse baselines. In figure 7(a), we show a comparison of our method (Dense VLAD SYNTH) to several baselines that use sparsely sampled local invariant features. For VLAD computed from (sparse) DoG keypoints, adding synthetic virtual views (Sparse VLAD SYNTH) helps (compared to Sparse VLAD). In contrast, adding synthetic virtual views to Fisher vector matching (Sparse FV SYNTH) does not improve over the standard FV without virtual views (Sparse FV). Overall, our method significantly improves over all sparse baselines.

Analysis of descriptor dimensionality. In figure 7(b) we investigate how the place recognition performance changes with reducing the dimensionality of the Dense VLAD descriptor from $4,096$ to $2,048$, $1,024$ and 512 dimensions. We observe a drop in performance specially for the lowest dimension. This suggests, that having a sufficiently rich representation is important for matching across large changes in appearance.

How many virtual views? In figure 7(c) we evaluate the required sampling of virtual views. First, we subsample the virtual views spatially from 5×5 meter grid (used in our method so far) to 10×10 meter grid. The spatial subsampling to 10×10 can reduce the number of virtual views by 75% with only a relatively small drop in place recognition performance. Then we subsample the number of yaw directions to only 6 per camera position, one every $60°$ (Dense VLAD SYNTH 60deg) compared to 12 yaw directions, one every $30°$ used in our method. In this experiment we keep the spatial sampling to 5×5 meters. Although the angular subsampling reduces the number synthetic views by only 50% it results in a fairly significant drop in performance, especially at the top 1 position.

Scalability. For the 24/7 Tokyo dataset, our method synthesizes $597,744$ virtual views compared to $75,984$ perspective street-view images in the same area. Hence, our method needs to index about 9 times more images compared to baselines without virtual view synthesis. We believe scaling-up towards place recognition in an entire city can be achieved with standard compression techniques such as Product Quantization (PQ) [20]. For example, the largest current place recognition benchmark by Chen *et al.* [8] that

Figure 9. **Example place recognition results with synthesized views (our method) compared to using only the original Google street-view images.** (Left) Query image. Note the difficult illumination. (2nd column) The best matching image (correct) by our method (Dense VLAD descriptor with the database expanded by synthesized views). (3rd column) The best matching image (incorrect) by Dense VLAD matching but using only the original street-view images. (4th column) The original street-view database image at the closest position to the query. Our method (2nd column) that uses virtual views with very similar viewpoints to the query can localize queries with difficult (night) illumination, thus enabling true 24/7 localization. This is not possible using the original street-view images (last column), which depict the same places but from quite different viewpoints. **Please see additional results on the project webpage [16]**

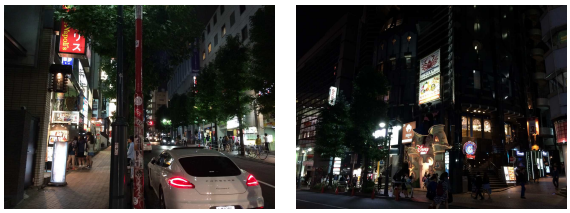


Figure 10. **Examples of challenging query images that remain hard to localize.**

covers a significant portion of the city of San Francisco contains 1M perspective images. We estimate that starting from a database of this size, but generating 9 times more virtual views with our SYNTH method, and compressing the resulting descriptors with PQ, would only require 2.9GB.

Qualitative results. Figures 8 and 9 show examples of place recognition results. Notice that query images (left column) include large changes in both viewpoint and illumination compared to the available street-view for the same places (right column). The synthesized views (2nd column) at new positions significantly reduce the variation in viewpoint and thus enable matching across large illumination changes, as discussed in section 3.

Limitations. Figure 10 shows examples of queries which remain very difficult to localize. The typical failure modes are (i) very dark night time images with limited dynamic range, (ii) places with vegetation, which is hard to uniquely describe using the current representation, and (iii) places where view synthesis fails often due to complex underlying 3D structure not captured well by the approximate depth maps available with street-view imagery.

6. Conclusion

We have described a place recognition approach combining synthesis of new virtual views with a densely sampled but compact image descriptor. The proposed method enables true 24/7 place recognition across major changes in scene illumination throughout the day and night. We have experimentally shown its benefits on a newly collected place recognition dataset – *24/7 Tokyo* – capturing the same locations in vastly different illuminations. Our work is another example in the recent trend showing benefits of 3D structure for visual recognition. As we build on the widely available Google street-view imagery our work opens-up the possibility of planet-scale 24/7 place recognition.

Acknowledgments. This work was partly supported by JSPS KAKENHI Grant Number 24700161, EU FP7-SPACE-2012-312377 PRoViDE, the ERC grant LEAP (no. 336845), ANR project Semapolis (ANR-13-CORD-0003) and the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

References

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, 2012.

[2] R. Arandjelović and A. Zisserman. All about VLAD. In *CVPR*, 2013.

[3] R. Arandjelović and A. Zisserman. DisLocation: Scalable descriptor distinctiveness for location recognition. In *Asian Conference on Computer Vision*, 2014.

[4] M. Aubry, B. C. Russell, and J. Sivic. Painting-to-3d model alignment via discriminative visual elements. *ACM Transactions on Graphics (TOG)*, 33(2):14, 2014.

[5] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[6] S. Cao and N. Snavely. Graph-Based Discriminative Learning for Location Recognition. In *CVPR*, 2013.

[7] S. Cao and N. Snavely. Minimal Scene Descriptions from Structure from Motion Models. In *CVPR*, 2014.

[8] D. Chen, G. Baatz, et al. City-scale landmark identification on mobile devices. In *CVPR*, 2011.

[9] 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 *ICCV*, 2007.

[10] M. Cummins and P. Newman. Highly scalable appearance-only SLAM - FAB-MAP 2.0. In *Proceedings of Robotics: Science and Systems*, Seattle, USA, June 2009.

[11] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, 2005.

[12] D. Hauagge and N. Snavely. Image matching using local symmetry features. In *CVPR*, 2012.

[13] D. Hauagge, S. Wehrwein, P. Upchurch, K. Bala, and N. Snavely. Reasoning about photo collections using models of outdoor illumination. In *BMVC*, 2014.

[14] J. Hays and A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.

[15] http://maps.google.com/help/maps/streetview/.

[16] http://www.ok.ctrl.titech.ac.jp/∼torii/project/247/.

[17] A. Irschara, C. Zach, J. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009.

[18] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *ICCV*, 2007.

[19] H. Jégou and O. Chum. Negative evidences and cooccurrences in image retrieval: the benefit of PCA and whitening. In *ECCV*, Firenze, Italy, 2012.

[20] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, 33(1):117–128, 2011.

[21] H. Jégou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

[22] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid. Aggregating local image descriptors into compact codes. *PAMI*, 34(9):1704–1716, 2012.

[23] B. Klingner, D. Martin, and J. Roseborough. Street view motion-from-structure-from-motion. In *ICCV*, 2013.

[24] J. Knopp, J. Sivic, and T. Pajdla. Avoding Confusing Features in Place Recognition. In *ECCV*, 2010.

[25] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Trans. Graphics*, 33(4), 2014.

[26] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[27] F. Li and J. Kosecka. Probabilistic location recognition using reduced feature set. In *Proc. Int. Conf. on Robotics and Automation*, 2006.

[28] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide Pose Estimation Using 3D Point Clouds. In *ECCV*, 2012.

[29] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[30] K. Matzen and N. Snavely. Scene chronology. In *ECCV*, 2014.

[31] M. Muja and D. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISAPP*, 2009.

[32] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[33] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007.

[34] T. Sattler, B. Leibe, and L. Kobbelt. Improving Image-Based Localization by Active Correspondence Search. In *ECCV*, 2012.

[35] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image Retrieval for Image-Based Localization Revisited. In *BMVC*, 2012.

[36] G. Schindler, M. Brown, and R. Szeliski. City-Scale Location Recognition. In *CVPR*, 2007.

[37] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz. Accurate geo-registration by ground-to-aerial image matching. In *3DV*, 2014.

[38] D. Sibbing, T. Sattler, B. Leibe, and L. Kobbelt. SIFT-Realistic Rendering. In *3DV*, 2013.

[39] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[40] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual Place Recognition with Repetitive Structures. In *CVPR*, 2013.

[41] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280, 2008.

[42] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[43] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys. 3D model matching with viewpoint-invariant patches (VIP). In *CVPR*, pages 1–8, June 2008.

[44] A. R. Zamir and M. Shah. Accurate Image Localization Based on Google Maps Street View. In *ECCV*, 2010.

[45] W. Zhao, H. Jégou, and G. Gravier. Oriented pooling for dense and non-dense rotation-invariant features. In *BMVC*, 2013.