C2SMART
CONNECTED CITIES WITH SMART TRANSPORTATION
A USDOT University Transportation Center

New York University
Rutgers University
University of Washington
University of Texas at El Paso
The City College of New York

Wearables to command more access and inclusion in a Smarter Transportation System

November 2022

C2SMART Center is a USDOT Tier 1 University Transportation Center taking on some of today's most pressing urban mobility challenges. Using cities as living laboratories, the center examines transportation problems and field tests novel solutions that draw on unprecedented recent advances in communication and smart technologies. Its research activities are focused on three key areas: Urban Mobility and Connected Citizens; Urban Analytics for Smart Cities; and Resilient, Secure, and Smart Transportation Infrastructure.

Some of the key areas C2SMART is focusing on include:

Disruptive Technologies

We are developing innovative solutions that focus on emerging disruptive technologies and their impacts on transportation systems. Our aim is to accelerate technology transfer from the research phase to the real world.

Unconventional Big Data Applications

C2SMART is working to make it possible to safely share data from field tests and non-traditional sensing technologies so that decision-makers can address a wide range of urban mobility problems with the best information available to them.

Impactful Engagement

The center aims to overcome institutional barriers to innovation and hear and meet the needs of city and state stakeholders, including government agencies, policy makers, the private sector, non-profit organizations, and entrepreneurs.

Forward-thinking Training and Development

As an academic institution, we are dedicated to training the workforce of tomorrow to deal with new mobility problems in ways that are not covered in existing transportation curricula.

Led by the New York University Tandon School of Engineering, C2SMART is a consortium of five leading research universities, including Rutgers University, University of Washington, University of Texas at El Paso, and The City College of New York.

c2smart.engineering.nyu.edu

Wearables to command more access and inclusion in a Smarter Transportation System

John Ross Rizzo
New York University
0000-0002-0366-8503

Chen Feng
New York University
0000-0003-3211-1576

Ruoyu Wang
New York University

Diwei Sheng
New York University

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

Acknowledgements

This research is funded by the Connected Cities for Smart Mobility towards Accessible and Resilient Transportation (C2SMART), a Tier 1 University Center awarded by U.S. Department of Transportation under contract 69A3351747124.

The authors would like to thank Carmera for providing the raw NYC image data set that we used for creating the NYU-VPR.

Executive Summary

Visual place recognition (VPR), technology often associated with navigation of autonomous vehicles, can be critical to meeting every day urban navigation needs of people with vision disabilities. This research addresses two major obstacles to implementing VPR at scale: 1) the need for side-view place recognition, crucial for identification of sidewalk features like storefronts; and 2) privacy concerns that result from capture of street-view images during the most relevant peak commute hours, and potential tension between obfuscation and inaccuracy that must be addressed before VPR database and query construction. Using an open-source dataset consisting of more than 200,000 images captured via camera-mounted taxis over a 2km by 2km area in Manhattan, New York, over the course of one year, researchers present benchmark results of the performance of popular VPR algorithms at both of these challenges. Results indicate that side-view recognition is significantly more challenging for current VPR methods, and that data anonymization has a negligible, or even marginally beneficial effect on performance.

This research contributes to the larger body of research in the following ways:

- Benchmarks VPR methods using a unique large-scale dataset of over 200,000 front-view and side-view images over a full year, capturing seasonal and other environmental variation
- Analyzes the causes of the significant challenges of VPR approaches using side-view images
- Using pixel removal as an anonymization technique and demonstrating that this anonymization has negligible impacts to VPR algorithm performance.

Table of Contents

List of Figures

List of Tables

Introduction

Introduction

Visual disabilities and impairment are associated with mobility losses, in addition to debility, illness and premature mortality. These mobility losses in people with moderate to severe visual disabilities are associated with an unemployment rate that approaches 60-80% in most developed countries.

Cities can provide a unique set of challenges for people with vision disabilities. In New York City, for example, and in similar cities around the United States, pedestrian crossing technology relies almost solely on visual “walk” and “don’t walk” cues connected to traffic management infrastructure, inaccessible to pedestrians who have difficulty seeing these cues. A recent 2021 New York City court ruling found that less than 5% of the 13,200 existing New York City signalized crosswalks included audible or tactile cues, found to be in violation of the Americans with Disabilities Act and the Rehabilitation Act, and as a result, 10,000 signalized crosswalks will be outfitted with Accessible Pedestrian Signal (APS) devices over the next years, with a judge-issued mandate that 100% of signalized crosswalks to have this APS technology installed by 2036.

Additional challenges predictably arise in other areas of urban navigation, such as entering and exiting subway stations and stores (hand-offs), and identifying store-fronts and other non-street urban features that are essential to day-to-day life for urban dwellers. This project is a continued partnership between Professor John-Ross Rizzo, Associate Professor of Neurology, Mechanical & Aerospace Engineering, and Biomedical Engineering, and Professor Chen Feng, Assistant Professor of Civil and Urban Engineering and Department of Mechanical and Aerospace Engineering at NYU Tandon School of Engineering. The proposed project would increase the safety profile and ease-of-use of **VIS^{4}ION** (**V**isually **I**mpaired **S**mart **S**ervice **S**ystem for **S**patial **I**ntelligence and **O**nboard **N**avigation), a

wearable personal mobility solution to assist people with vision disabilities with urban navigation. The product serves as a customizable, human-in-the-loop, sensing-to-feedback platform to deliver navigation assistance.

The objective of the project was to enhance the platform by advancing the mapping and localization software through the creation of ethograms from user studies that would characterize the mobility behavior of pedestrians with visual disabilities, resulting in improved wayfinding through sidewalks, intersections, trains, and bus stops. Integration of these ethograms into the VIS^{4}ION platform would enhance its ability to assist with localization, mapping, orientation and direction of users in urban areas.

Three changes over the course of 2020 resulted in significant changes to the project scope. The first was the impact of the COVID-19 pandemic, especially in New York City in which the research team is based. As researchers were seeking to understand the severity, contagion, and transmission mechanisms of COVID-19, Institutional Research Boards (IRB) across universities enacted a blanket suspension of all in-person research activity. This suspension prevented the recruitment of pedestrians who are blind or have moderate to severe visual disabilities for user tests, and threatened to derail the project.

This research project was re-scoped to emphasize a second important thrust of enhancements to the VIS4ION platform: computer vision-aided localization refinement. The flow of the proposed

navigation method is illustrated in **Figure 1** and contains two phases: mapping, and localization.

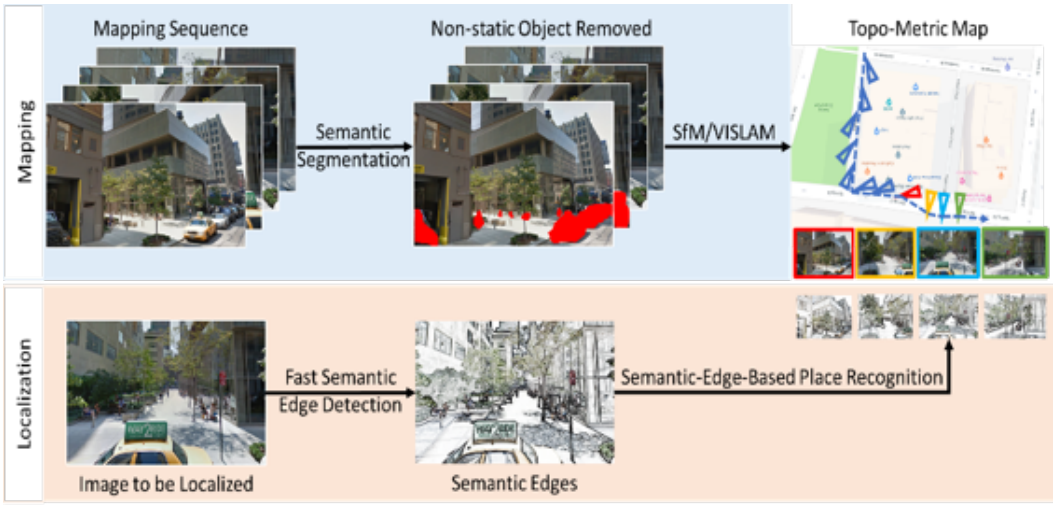


Figure 1. Pipeline of the proposed navigation system, separated into a mapping sequence (Top, blue panel) and a localization sequence (Bottom, red panel).

During the mapping phase, multiple videos are captured on a target area. These videos are processed to create a map representing the geometry and appearance of an area through a semantic 3D reconstruction neural net, Map-Net. In the localization phase, an image captured in-real time (e.g. from a wearable navigation assistant) is correlated with the mapping data to produce detailed navigation information. An example is shown in Figure 1, in which a 3D map is created of an intersection in Brooklyn, New York, using a sequence of captured images, and then trained on an image-based place recognition system.

In light of IRB suspension on human subject research, the research team rescoped the project to focus on refining computer vision-aided mapping and localization using footage captured by vehicles rather than footage captured by users equipped with VIS4ION or other navigation assistants.

This research inquiry was aided by collaboration with Professor Claudio Silva, Institute Professor of Computer Science and Engineering at NYU Tandon School of Engineering, and Professor of Data Science at the NYU Center for Data Science. Researchers led by Professor Silva were collaborating with

Carmera (later acquired by WovenPlanet, a Toyota subsidiary) and presented a seminal dataset of urban images captured by cameras mounted on taxis. This large scale urban dataset provided a means to evaluate the performance of leading visual place recognition (VPR) algorithms to with the larger goal of improving assistive navigation for people with visual disabilities, especially urban areas.

VPR refers to matching images with portions of similar images queried from a large database of images capture by known camera positions. Its assistive navigation applications range from autonomous driving to pedestrian navigation and is especially promising in "urban canyons," dense and complex urban areas, such as New York City, in which GPS precision is insufficient for real-time navigation needs or satellite signals are reflected or obfuscated leading to localization errors.

Large-scale image capture via car dashboard-mounted camera is mostly commonly performed with a focus on the "front view," or the direction parallel to the car's driving direction, capturing the street and sky immediately in front of the car—most relevant to a car's navigation. This front-view, the default choice in VPR focused on autonomous vehicle applications, is often less relevant to pedestrians navigating a city street, in which daily transportation activity may involve locating restaurants and store-fronts and entering and existing metro stations; front-view images consist of upwards of 50% of pixels that are not relevant to some wearable navigation technology which rely on fronto-parallel images of buildings to provide relevant information.

However, there has been a historical lack of research focus into datasets which consist of side-view images in large enough, isolated quantities to conduct rigorous comparisons. To the PIs' knowledge at the time of this report, there had not been a systematic performance comparison between front and side-view images.

Dataset	Side-view	Side-view label	Dynamic-Object	Crowded-Area	Anonymization	Seasonal-changes	#images

StreetLearn	✓	-	✓	✓	×	×	143,000
StreetView	✓	-	✓	✓	×	×	62,058
Nordland	×	-	×	×	×	✓	28,865
VPRiCE 2015	×	-		×	×	×	7,778
Tokyo 24/7	✓	×	✓	✓	face-only	×	76,000
Pittsburgh	✓	×	✓	×	×	✓	254,064
KITTI raw	×	-	✓	×	×	×	12,919
KAIST	×	-	✓	×	×	×	105,000
Oxford RobotCar	×	-	✓	×	×	✓	19,556,490
Mapillary	✓	✓	✓	✓	×	✓	1,681,000
NCLT	×	-	✓	×	×	✓	100,000
NYU-VPR (ours)	✓	✓	✓	✓	✓	✓	201,790

Table 1: Comparison of Major Public Outdoor VPR Datasets with NYU-VPR

The open-source dataset used in this research project consists of more than 200,000 images. The dataset is being used to test different VPR approaches. In particular, this research has advanced an understanding of side-view versus front-view recognition, which are essential to enable pedestrian navigation and interaction with shops, metro stations and other features of the pedestrian urban environment.

This research inquiry seeks to inform the following questions: do side-view images present an increased performance challenge to VPR methods than do front-view images? If so, what is the magnitude of this challenge, and why?

In addition, this research seeks to understand the impact of image anonymization on VPR performance. Because large scale datasets of images contain personally identifiable information of pedestrians and license plates and are stored over extended periods of time, images must be anonymized. In particular, this research addresses anonymization by wiping all identity-related pixels and seeks to understand how, and to what extent, this anonymization method might affect accuracy and robustness of VPR algorithms.

Side-View Challenges

There are two main reasons that side-view images were hypothesized to pose more of a challenge to VPR methods: 1) there is much less of an overlap between two sequential-side-view images than two sequential front-view images. The size of overlap between consecutive side-images is even greater if storefronts are closer to the camera (smaller field of vision) or on narrow streets. 2) The presence of trees and scaffolding, visually similar to each other even in completely different images, may impact localization accuracy. 3) Motion blur poses a more serious problem for capturing side-view images, which are smaller in size and therefore relatively more affected, than for front-view images.

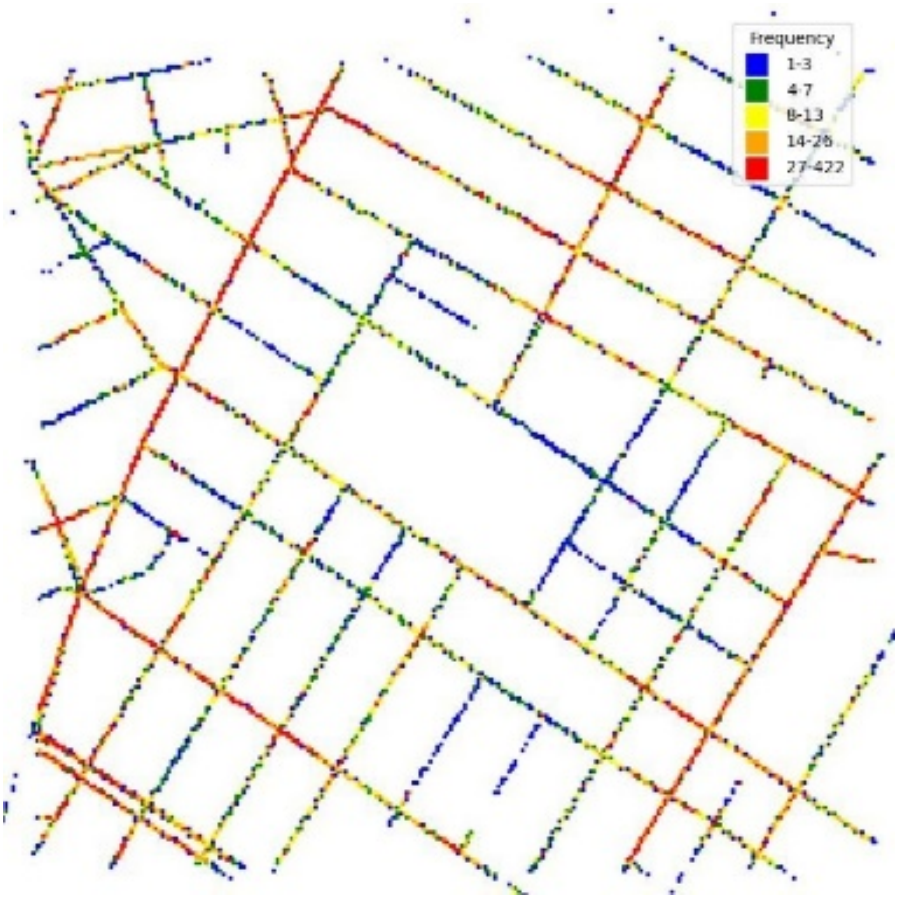


Figure 2. Frequency of capture location

More than 200,000 images taken over the course of the year April 2016 to March 2017 in 2km by 2km area around the Washington Square Park area in Manhattan, New York. Images were taken from smart-phone cameras mounted on the front, back and sides of undisclosed taxis, which randomized the frequency of captured images, using auto-exposure, and tagged with GPS. The full dataset consists of the following images:

- 100,500 side-view
- 101,290 front-view
- 640 x 480 resolution

This dataset is unique in that it compares front-view images, which capture sky and road features, with side-view images, which capture store-fronts, subway entrances, and shop signs to assist with 360 urban navigation. This captures all 4 seasons, and resulting changes in season like snow and heavy Fall foliage.

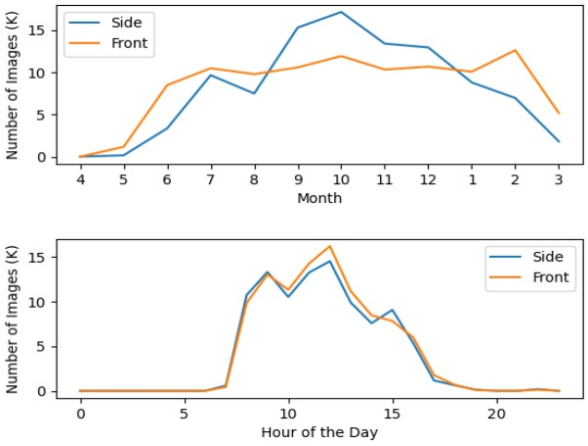


Figure 3. Distribution of side and front view images over time

It also captures changes in the urban landscape like construction and street closures, and anonymizes pedestrians and license plates.

Methods

Difficulty Level

Side-views were assigned a difficulty level in the following ways:

Scale Invariant Feature Transform (SIFT) features were extracted for each query image

The top-8 closest images to each query images were identified by GPS coordinates; the query image and this top-8 form 8 image pairs. Random sample consensus (RANSAC) was used to compute a fundamental matrix to identify the number of inliers for each image pair. The difficulty level of matching each image with pair was assigned an interval based on each pair's number of inlier points: 0-19 (hard), 20-80 (medium), and greater than 80 (easy). The intervals were designated based on the similarity of each image pair, with the intervals for side-view images based on the most common difficulty of each of its 8

image pairs. Multi-domain Semantic Segmentation (MSEG) was used to anonymize the dataset by replacing people and cars with white pixels.

Figure 4. Raw vs Anonymized Images

VPR Methods

VPR methods can roughly be grouped into three categories: deep-learning-based, non-deep-learning-based, and methods that use only deep-learning-based descriptors. All three categories are explored in this research. Deep-learning-based methods use convolutional neural networks (CNN) trained in an end-to-end manner. Non-deep-learning methods include bag-of-words (BOW) models, and Vector of Locally Aggregated Descriptors (VLAD). Researchers used DBoW+ORB, used in the popular ORB-SLAM for loop closing, and VLAD+SURF (speeded up robust features.) Deep-learning-based descriptors methods rely on deep nets' detection of a richer set of key points, such as SuperPoint, which was also adopted for benchmarking.

The dataset was randomly divided into training (80% of images,) validation (5% of images) and testing (15% of images) groups. The Python module utm was used to convert GPS coordinates to Universal Transverse Mercator (UTM) coordinates to increase the precision of distance calculations. The following methods were then evaluated in this research:

- Vector of Locally Aggregated Descriptors(VLAD) with speeded up robust features (SURF)

- VLAD with SuperPoint pre-trained on the MS-COCO (Microsoft Common Objects in Context) image dataset, containing 328,000 images of common objects and humans used to train machine learning models.
- NetVLAD
- PoseNET
- Distributed Bag of Words (DBoW)

VLAD+SURF

SURF descriptors were aggregated for image retrieval using VLAD. Researchers used the MiniBatchKmeans algorithm with batch size set to 5,000 to determine an optimal cluster number of 32 within 8, 16, 32, and 64, resulting in high accuracy and acceptable training time: 8 hours to train 77,608 images on a CPU with 64GB of available memory.

VLAD+SuperPoint

Using a SuperPoint model pre-trained on an MS-COCO generic image dataset, researchers extracted SuperPoint features using nVidia RTX 2080S. SuperPoint descriptors were aggregated for image retrieval using VLAD, again with cluster number set at 32 after using the MiniBatchKmeans algorithm with batch size set to 100. The SuperPoint descriptors are much larger than SURF descriptors, and it took 20 hours to train the 77,608 training image set on a CPU and GPU with 64 GB of available memory.

NetVLAD

The pre-trained model weight was used for 30 epochs on the Google Street View Pittsburgh-250k dataset. CPU used was the Intel Core i7-8700k; NVIDIA GEFORCE GTX 1080 TI was used for GPU. Initial

clustering was conducted on the training data to determine centroids used for testing. Then, the input testing data with extracted deep feature are assigned to clusters, using batch size of 24.

PoseNet

PoseNet model was used with ResNet34 as base architecture. Cartesian coordinates of images are required as inputs for training PoseNet, so latitude and longitude of training images were gathered and converted to universal transverse Mercator (UTM) coordinates. These normalized UTM coordinates were used as inputs, and improve the accuracy of PoseNet's estimation of image relative position.

DBoW

For the Distributed Bag of Words (DBoW) model, researchers chose Oriented FAST and Rotated BRIEF (ORB) descriptors to represent features. DBoW was used to generate a vocabulary constructed by ORB descriptors of training and test images. The top-5 retrieved images were identified by using DBoW to generate a score between each training and test image and identifying the top-5 scores for each test image.

Discussion

Results

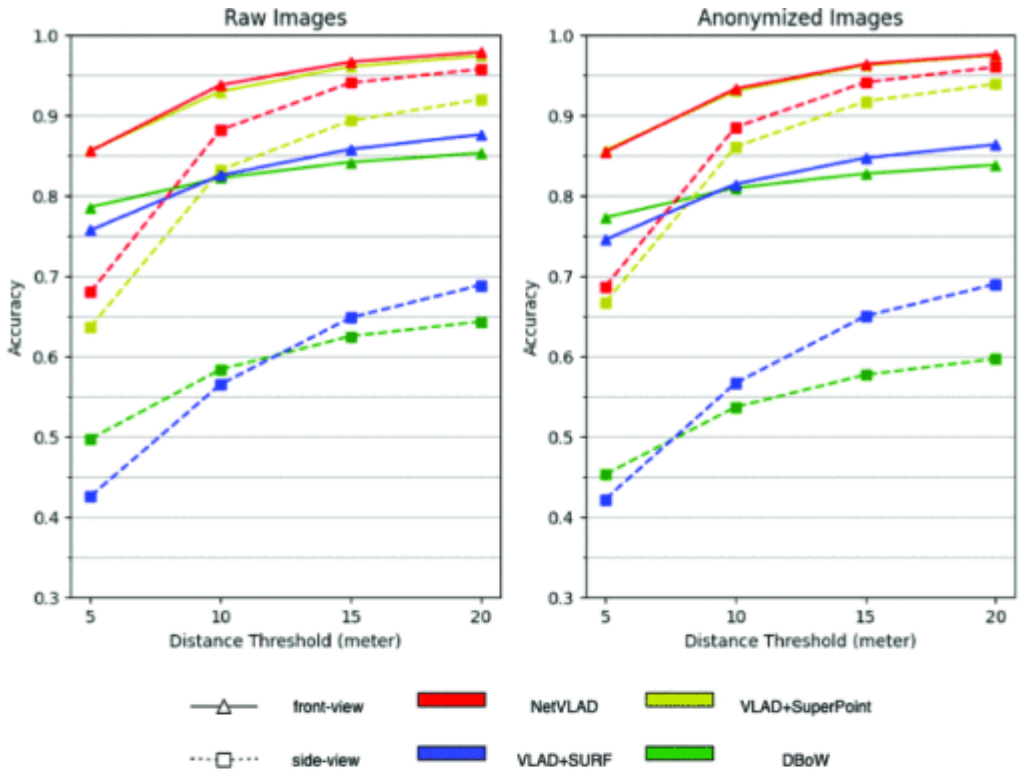
Evaluation

Top-1 and top-5 retrieval accuracy was measured using four distance thresholds: 5, 10, 15, and 20 meters. If any of the top-k retrieval images are within the given distance threshold from the query image, it is counted as a successful retrieval. The top-k ranking is based on the similarity between image features calculated by VPR algorithms. This evaluation metric is similar to the more commonly used precision-recall curve.

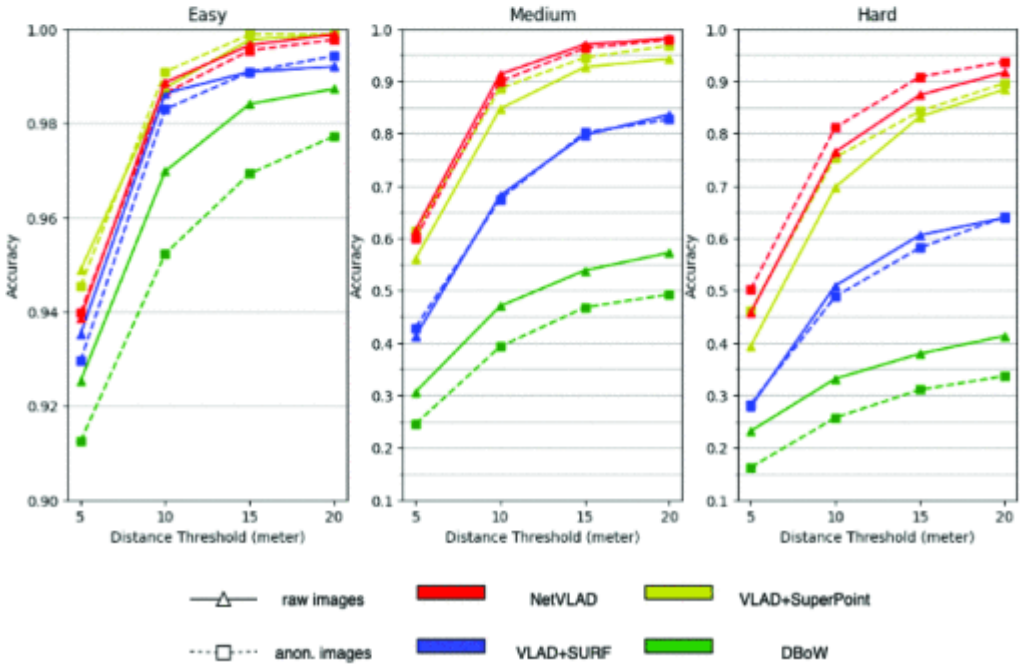
Performance

Predictably, top 5 retrieval image accuracy is higher than top 1 retrieval image accuracy by 10% on average. Use of VLAD to aggregate descriptors results in SuperPoint descriptor accuracy greater than SURF descriptors. NetVLAD is most accurate, followed by VLAD and SuperPoint, followed by VLAD and SURF, followed by the DBoW. The low DBoW accuracy is attributed to the unsustainability of ORB features.

PoseNet outputs a GPS coordinate, which can be input to find the closest top 1 retrieval image. Through experiments, accuracy of PoseNet is 15.3% when the distance threshold is 5 meters, and 37.5% when the distance threshold is 10 meters. Due to its low performance, PoseNet was omitted from subsequent experiments, and results were not plotted. Accuracy was plotted by difficulty level in Figure 5.



(a) Top-5 retrieval results.



(b) Top-5 retrieval results in terms of difficulty level.

Figure 5. Accuracy plotted by difficulty level

Anonymization:

Anonymization does not appear to have a large impact on VPR results for either front or side-view data, resulting in a decrease of 2.1% and 3.4% for DBoW and VLAD+SURF, resepectively. Interestingly, anonymization increases accuracy, by 1.1% on average for VLAD with SuperPoint. This result suggests that anonymization for privacy purposes will not significantly affect VPR experiments.

View Direction

Camera view direction does appear to have a large impact on VPR results. Front-view images are more accurate than are side-view images, across all VPR methods, and anonymization reduces side-view accuracy while it increases accuracy front-view images. The PIs hypothesize that there are more street features blocked by anonymized pedestrians and cars in side-view images, which makes the image more difficult to recognize. Front-view images, however, appear to benefit from a reduction in image noise, resulting in accuracy improvements.

Challenges

Major challenges to this research methodology include changes in seasons, changes in construction environments, and changes in speed resulting in blurry images. Because images were captured over the course of a year, images taken of the same locations will appear different due to seasonal variation like the presence of snow or leaves, and changes in street construction (like the presence or absence of

scaffolding.)

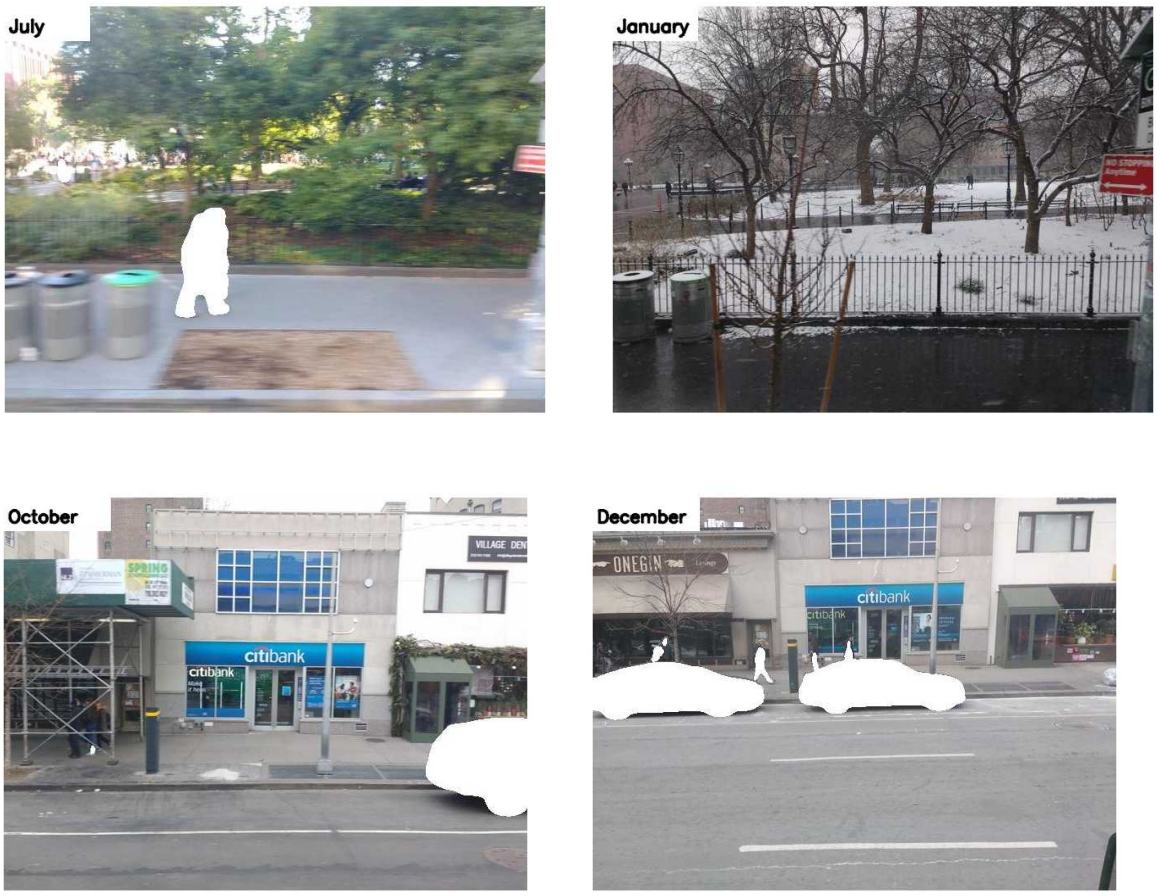


Figure 7. Images during and after construction

Because images were captured from taxi mounted cameras, there are instances of blurry images due to increasing speed during low-traffic intervals.

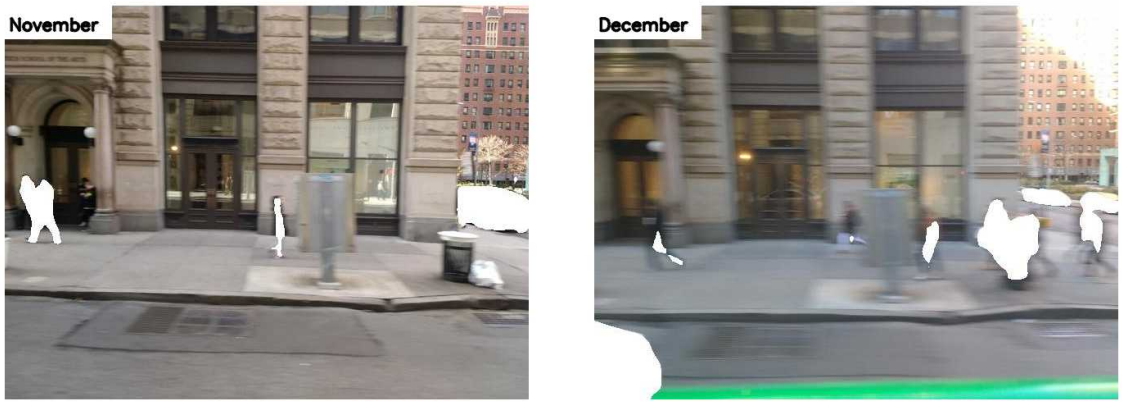


Figure 8. Images captured with and without motion blur

Outputs

This research produced an open-source dataset with more than 200,000 front-view and side-view outdoor images taken in a 2km by 2km area around the Washington Square Park area in Manhattan, New York. This data, and the benchmark code, are released for educational and research purposes.

In addition, the research team presented this research at 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

Conclusion

Despite setbacks caused by the COVID-19 pandemic and the ensuing IRB prohibition on user testing, advancement was made in the field of visual place recognition (VPR), central to improving assistive navigation for people with visual disabilities, especially in urban areas, in collaboration with Professor Claudio Silva, and a large raw NYC image dataset captured by Carmera. Specifically, this research evaluates a unique large-scale year-long image dataset used to evaluate the performance of popular VPR algorithms. This research finds that side-view images present a larger challenge to VPR methods than do front-view images, with significant reduction in performance across all VPR methods tested. In addition, data anonymization does not significantly affect VPR algorithm performance. Moreover, marginal improvements in VPR performance were observed on anonymized images, potentially due to removal of noise.

Future work will benchmark additional VPR methods with the end goal of improving ease of urban navigation for people with visual disabilities, and especially reducing the impact on the unemployment rate attributed to mobility challenges.

References

1. P. Mirowski, A. Banki-Horvath, K. Anderson, D. Teplyashin, K. M. Hermann, M. Malinowski, M. K. Grimes, K. Simonyan, K. Kavukcuoglu, A. Zisserman et al., “The streetlearn environment and dataset,” arXiv preprint arXiv:1903.01292, 2019. 2
2. A. R. Zamir and M. Shah, “Image geo-localization based on multiplenearest neighbor feature matching using generalized graphs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 8, pp. 1546–1558, 2014. 2
3. N. S¨underhauf, P. Neubert, and P. Protzel, “Are we there yet? challenging seqslam on a 3000 km journey across all four seasons,” in Proc. IEEE Int’l Conf. Robotics and Automation (ICRA), 2013, p. 2013. 2
4. “The VPRiCE Challenge 2015 – Visual Place Recognition in Changing Environments - Public - Confluence.” [Online]. Available: https://roboticvision:atlassian:net/wiki/spaces/PUB/pages/14188617/The+VPRiCE+Challenge+2015+Visual+Place+Recognition+in+Changing+Environments 2
5. A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, “24/7 place recognition by view synthesis,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1808–1817. 2, 3
6. A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, “Visual place recognition with repetitive structures,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2013. 2, 4
7. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” Int’l J. Robotics Research, 2013. 2

8. Y. Choi, N. Kim, K. Park, S. Hwang, J. S. Yoon, Y. In, and I. Kweon, “All-day visual place recognition: Benchmark dataset and baseline,” in Workshop on Visual Place Recognition in Changing Environments, 06 2015. 2
9. W. Maddern, G. Pascoe, C. Linegar, and P. Newman, “1 year, 1000 km: The oxford robotcar dataset,” Int’l J. Robotics Research, vol. 36, no. 1, pp. 3–15, 2017. 2
10. F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, “Mapillary street-level sequences: A dataset for lifelong place recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 2626–2635. 2, 3
11. N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, “University of michigan north campus long-term vision and lidar dataset,” The International Journal of Robotics Research, vol. 35, no. 9, pp. 1023–1035, 2016. 2, 3
12. P. Speciale, J. L. Schonberger, S. B. Kang, S. N. Sinha, and M. Pollefeys, “Privacy preserving image-based localization,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 5493–5503. 1
13. J. Lambert, Z. Liu, O. Sener, J. Hays, and V. Koltun, “MSeg: A composite dataset for multi-domain semantic segmentation,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2020. 2, 3
14. R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5297–5307. 3
15. A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocalization,” in Proc. IEEE Int’l Conf. Computer Vision (ICCV), 2015, pp. 2938–2946. 3, 4

16. M. Chanc´an, L. Hernandez-Nunez, A. Narendra, A. B. Barron, and M. Milford, “A hybrid compact neural architecture for visual place recognition,” IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 993–1000, April 2020.
17. Z. Chen, A. Jacobson, N. S¨underhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, “Deep learning features at scale for visual place recognition,” in Proc. IEEE Int’l Conf. Robotics and Automation (ICRA). IEEE, 2017, pp. 3223–3230. 3
18. D. G´alvez-L´opez and J. D. Tard´os, “Bags of binary words for fast place recognition in image sequences,” IEEE Trans. R obotics, vol. 28, no. 5, pp. 1188–1197, October 2012. 3
19. H. J´egou, M. Douze, C. Schmid, and P. P´erez, “Aggregating local descriptors into a compact image representation,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR). IEEE, 2010, pp. 3304–3311. 3, 4
20. T. Sattler, B. Leibe, and L. Kobbelt, “Improving image based localization by active correspondence search,” in Proc. European Conf. Computer Vision (ECCV). Springer, 2012, pp. 752–765. 3
21. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in Proc. IEEE Int’l Conf. Computer Vision (ICCV), 2011, pp. 2564–2571. 3, 5
22. R. Mur-Artal, J. M. M. Montiel, and J. D. Tard´os, “Orbslam: A versatile and accurate monocular slam system,” IEEE Trans. Robotics, vol. 31, no. 5, pp. 1147–1163, 2015. 3
23. H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in Proc. European Conf. Computer Vision (ECCV). Springer, 2006, pp. 404–417. 3, 4
24. X. Yu, S. Chaturvedi, C. Feng, Y. Taguchi, T.-Y. Lee, C. Fernandes, and S. Ramalingam, “VLASE: Vehicle localization by aggregating semantic edges,” Proc. IEEE/RSJ Int’l Conf. Intelligent Robots and Systems (IROS), 2018. 3, 5

25. D. DeTone, T. Malisiewicz, and A. Rabinovich, “Superpoint: Self-supervised interest point detection and description,” in CVPR Deep Learning for Visual SLAM Workshop, 2018. [Online]. Available: http://arxiv:org/abs/1712:07629 3, 4
26. D. G. Lowe, “Object recognition from local scalein variant features,” in Proc. IEEE Int’l Conf. Computer Vision (ICCV), vol. 2, 1999, pp. 1150–1157 vol.2. 3