2023 IEEE International Conference on Robotics and Automation (ICRA)
Workshop on Active Methods in Autonomous Navigation
29 May – 2 June, 2023, London, UK

Visual Place Recognition with Low-Resolution Images

Mihnea-Alexandru Tomiță[1], Bruno Ferrarini[1], Michael Milford[2], Klaus McDonald-Maier[1], Shoaib Ehsan[1,3]

***Abstract*— Images incorporate a wealth of information from a robot's surroundings. With the widespread availability of compact cameras, visual information has become increasingly popular for addressing the localisation problem, which is then termed as Visual Place Recognition (VPR). While many applications use high-resolution cameras and high-end systems to achieve optimal place-matching performance, low-end commercial systems face limitations due to resource constraints and relatively low-resolution and low-quality cameras. In this paper, we analyse the effects of image resolution on the accuracy and robustness of well-established handcrafted VPR pipelines. Handcrafted designs have low computational demands and can adapt to flexible image resolutions, making them a suitable approach to scale to any image source and to operate under resource limitations. This paper aims to help academic researchers and companies in the hardware and software industry co-design VPR solutions and expand the use of VPR algorithms in commercial products.**

***Index Terms*— Visual Place Recognition, Visual Localisation**

I. Introduction

With the advances in technology made in the last decade, image and video capturing devices became exceptional in reproducing a higher quality representation of our surroundings. Visual Place Recognition (VPR) utilises the visual information gathered from the camera to perform the localisation process. To achieve high place matching performance, VPR applications usually employ high-end systems and advanced cameras [1]. However, low-end commercial products are computationally limited and have low-resolution cameras. Thus, the deployment of robust but computationally demanding VPR methods is restricted on such platforms, as identified in [2], [3]. Hence, handcrafted VPR techniques are suitable to be deployed on resource constrained platforms, due to their computationally efficient nature. In addition to their low computational requirements, handcrafted VPR techniques can adapt to various image resolutions, which makes them attractive for VPR applications on resource-constrained platforms, with low-resolution cameras. Moreover, as a lower-resolution image is visually different from its high-resolution version (refer to Fig. 1), this paper analyses the optimal image resolution for different handcrafted descriptors. Thus, the aim of this paper is to reduce the image resolution to facilitate VPR applications on resource-constrained commercial platforms. In summary, our contributions are as follows:

- An assessment of the performance of several well-established handcrafted VPR techniques on various image resolutions. We employ several datasets to enable a VPR performance comparison in real-world scenarios, under illumination, viewpoint and seasonal variation.
- We report the total time required to perform VPR for each descriptor, showing how a reduced image resolution results in a more efficient VPR process. We also perform a trade-off analysis between performance and computation, showing the best descriptor that should be selected depending on the image resolution.

The remainder of this paper is organised as follows: Section II presents the literature review. Section III presents the experimental setup, where the VPR time is discussed. We also present the performance metric employed, together with the selection of VPR techniques and datasets utilised in this study. Section IV presents the detailed results and analysis. The conclusions are presented in section V.

Fig. 1. The same image resized to various resolutions.

TABLE I
THE SIZE OF EACH DATASET IN MEGABYTES (MB) RESIZED TO VARIOUS RESOLUTIONS.

Dataset	Image Resolution [px]						
	16x16	32x32	64x64	128x128	256x256	512x512	1024x1024
17 places	0.671	0.872	1.5	3.3	8.4	21.8	57.9
Campus Loop	0.151	0.194	0.339	0.845	2.7	9.3	28.7
Gardens Point	0.442	0.573	1	2.5	7.5	23.1	63.9
Nordland	0.25	0.311	0.512	1.2	3.4	10.1	29.8
SYNTHIA	0.296	0.379	0.647	1.5	4.6	16.2	56.9

II. Literature Review

Prior to the deployment of deep-learning for VPR applications, handcrafted local feature descriptors were primarily utilised to solve VPR challenges. However, these cannot handle severe illumination changes in the environment. In contrast with local feature descriptors that analyse keypoints

[1]Authors are with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ, United Kingdom matomi@essex.ac.uk, bferra@essex.ac.uk, kdm@essex.ac.uk, sehsan@essex.ac.uk.

[2]Michael Milford is with the School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia michael.milford@qut.edu.au

[3]Shoaib Ehsan is also with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom s.ehsan@soton.ac.uk

This work is supported by the UK Engineering and Physical Sciences Research Council through grants EP/R02572X/1 and EP/P017487/1.

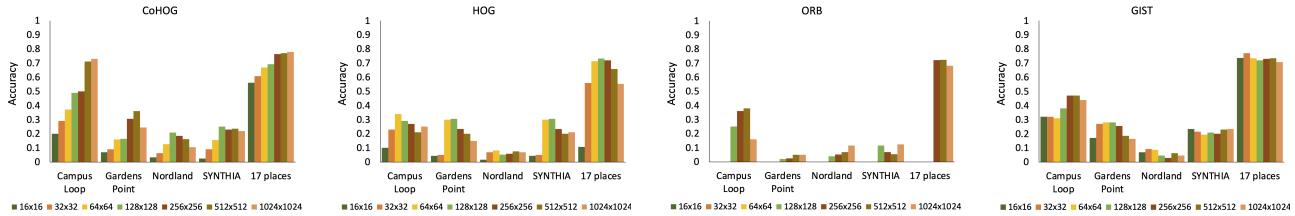

Fig. 2. The accuracy of all VPR techniques on each resized dataset.

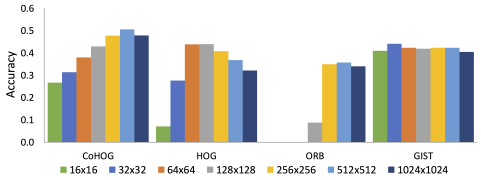

Fig. 3. The average accuracy of each technique on the combined datasets.

Fig. 4. Keypoints found in the same image at several distinct resolutions, as determined by ORB descriptor.

in images, global feature descriptors analyse the entire content of the image [1].

A popular global feature descriptor is GIST [4] utilised in VPR applications such as [5], [6]. HOG [7] is a computationally efficient global descriptor, tolerant to appearance changes [8]. In [9], the authors propose CoHOG, a compute-efficient, and training-free VPR system based on HOG descriptors, having good tolerance to lateral shifts.

Local feature descriptors such as SURF [10] and SIFT [11], [12] have been widely used in VPR applications such as [13], [14]. BRIEF [15] has similar VPR performance with SIFT and SURF, albeit at a reduced encoding time. Bag-of-Words model (BoW) [16] and Vector of Locally Aggregated Descriptors (VLAD) [17] build an image descriptor of fixed length by aggregating local feature descriptors around centroids. BoW and VLAD can be used for VPR as shown in [17] and [3], respectively. The authors of ORB [18] propose a computationally-efficient descriptor, capable of performing real-time VPR.

In this work, the focus is mainly towards global feature descriptors, as local features descriptors are unable to detect keypoints in small images, as later shown in section IV-A. Thus, they are not suitable to operate on small resolution images.

III. Experimental Setup

A. VPR Time

For low-end commercial products which are computationally limited, it is important to determine the optimal technique. Hence, in this paper we utilise the total time required to perform VPR (t_{VPR}) as a measurement of computational efficiency. The t_{VPR} of each technique is determined by adding the encoding time t_e with the matching time t_m as follows:

$$t_{VPR} = t_e + t_m, \tag{1}$$

where t_e refers to the amount of time that a VPR technique requires to compute the feature descriptor of an image and t_m represents the time required to match the descriptor of a query image with all the reference descriptors in the map.

TABLE II

The encoding time in milliseconds (ms) of a query image, for each VPR technique.

VPR Technique	Image Resolution [px]						
	16x16	**32x32**	**64x64**	**128x128**	**256x256**	**512x512**	**1024x1024**
CoHOG	14.5	15.3	16.6	19	30.6	77.3	260.1
HOG	0.104	0.236	1.307	0.514	1.585	6.578	31.32
ORB	-	-	-	0.86	2.49	6.171	17.5
GIST	0.967	2	9.807	27.561	153.99	708.02	4618.1

TABLE III

The average matching time in milliseconds (ms), for each VPR technique.

VPR Technique	Image Resolution [px]						
	16x16	**32x32**	**64x64**	**128x128**	**256x256**	**512x512**	**1024x1024**
CoHOG	2.806	2.985	3.634	11.9	81.25	852.12	15398.19
HOG	2.887	3.036	3.166	3.749	6.318	15.79	50.864
ORB	-	-	-	30.16	263.12	400.7	392.4
GIST	2.32	2.247	2.375	2.309	2.389	2.418	2.67

B. Performance Metric

To evaluate the VPR performance on various image resolutions, the percentage of correctly matched images is utilised, having the following formula:

$$Accuracy = \frac{N_c}{N_q}, \tag{2}$$

where N_c represents the number of correctly matched query images and N_q is the total number of query images. The accuracy has values in range [0,1], higher values denoting better VPR performance.

C. VPR Techniques

A selection of four well-established VPR techniques have been employed in this work including: HOG [7], CoHOG [9], ORB [18] and GIST [4]. For HOG, a cell and block size of 16x16 pixels was utilised, with a total number of histogram bins of 9 [8]. The remaining VPR techniques have

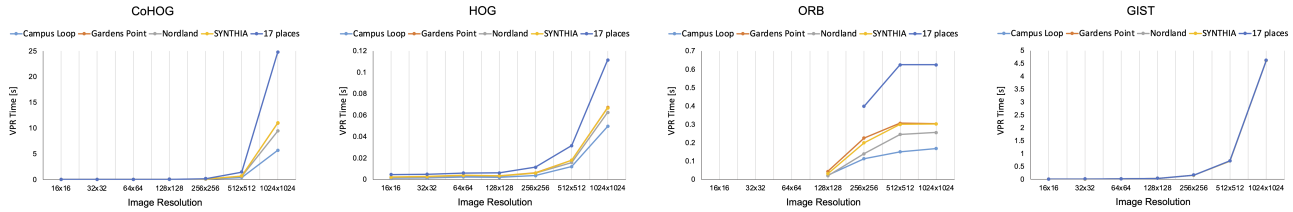

Fig. 5. The VPR time (refer to equation (1)) in seconds (s) of all VPR techniques on various image resolutions.

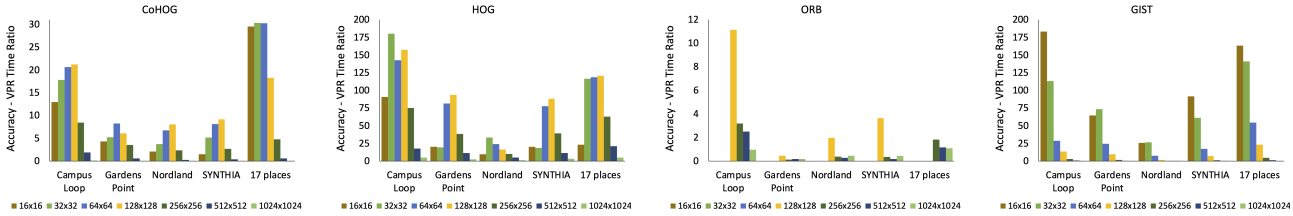

Fig. 6. The ratio between accuracy and VPR time of each technique on all resized dataset.

been utilised as presented by their authors, with no additional changes being made to neither technique.

D. Test Datasets

In this paper, five well-established VPR datasets are employed to present our findings. These are as follows: Campus Loop dataset [19] consists of 100 query and 100 reference images, with a large amount of frames that contain viewpoint and seasonal variations; Gardens Point dataset [20] consists of 200 query (*day_left*) and 200 reference (*night_right*) images, with a focus on illumination and viewpoint variation; Nordland dataset [21] captures the drastic changes between seasons. In this paper, we have utilised 172 query images (summer) and 172 reference images (winter) of the Nordland dataset. SYNTHIA dataset [22] is a simulated city-like environment that consists of 200 query and 200 reference images, taken in various weather, seasonal and illumination conditions. 17 places [23] is an indoor dataset, whose images contain illumination and viewpoint variation. For this study, three locations have been selected entitled Arena, AshRoom and Corridor. Hence, this dataset consists of 457 query (*day_vme1*) and 434 reference images (*night_vme1*).

To enable a place matching performance comparison of each technique employed, the above mentioned datasets have been resized to several image resolutions (values presented in pixels (px)), ranging from 16x16 px to 1024x1024 px. Fig. 1 presents some sample images taken from the Gardens Point *day_left* dataset resized to various image resolutions. Table I presents the size in Megabytes of each resized dataset.

IV. Results and Analysis

A. Place Matching Performance

The performance of all VPR techniques on every resized dataset is presented in Fig. 2. In contrast with the VPR accuracy of HOG and GIST which peaks towards smaller images, CoHOG benefits from an increased image resolution. Moreover, as CoHOG is designed to handle lateral shifts in camera movement, this technique achieves high accuracy on 17 places and Campus Loop datasets, while utilising a higher image resolution than the rest of the techniques (1024x1024 px). This trend is also emphasized in Fig. 3, which presents the average performance for each technique on all presented datasets, where the accuracy for each image resolution is weighted with regards to the number of images in the dataset. CoHOG achieves the highest place matching performance on datasets resized to 512x512 px. For GIST, the highest accuracy is reported on the datasets resized to 32x32 px. HOG achieves similar levels of performance on both 64x64 px and 128x128 px resized datasets, as seen in Fig. 3. It is important to note that ORB cannot work with small image resolutions, as previously mentioned in section II. This happens because no keypoints are detected in images, or the image is smaller than the descriptor patch. In our experiments, ORB cannot work with image resolutions of less than 128x128 px. Moreover, for 17 places dataset, ORB does not find any keypoints in image resolutions of less than 256x256 px. Fig. 4 shows the keypoint locations of ORB at different image resolutions. It can be seen that by reducing the image resolution to 64x64 px, ORB fails to detect any of the previously identified keypoints in the presented environment.

B. Analysis on the Time Required to Perform VPR

This sub-section performs an analysis on the total time required to perform VPR. Table II presents the encoding time t_e and Table III presents the matching time t_m of all VPR techniques. Moreover, the VPR time (refer to equation (1)) of every technique is presented in Fig. 5. We have previously discussed in sub-section IV-A that CoHOG achieves increased levels of place matching performance utilising a higher image resolution. However, as the matching time t_m of CoHOG is really high when utilising a higher image resolution, its VPR time is drastically increased, as seen in Fig. 5 on the 17 places dataset. When utilising an image resolution of 128x128 px and above, GIST achieves high encoding times when compared to the remaining VPR tech-

niques, as reported in Table II. In contrast with HOG, ORB and CoHOG where the descriptor size changes depending on the image resolution, the descriptor size of GIST always remains constant, therefore its matching time (refer to Table III) remains similar for every image resolution. Thus, the t_{VPR} of GIST does not differ significantly from one dataset to another, as shown in Fig. 5. Hence, GIST should be selected for VPR applications with a focus on fast processing times. However, if the aim is towards VPR performance, CoHOG should be utilised instead.

C. Performance and Computation Trade-off Analysis

As utilising a lower image resolution generally results in a decrease in t_{VPR} (refer to Fig. 5), this section performs a trade-off analysis between VPR performance and time. The ratio between the accuracy and VPR time of each technique on all resized datasets is presented in Fig. 6. As previously mentioned in sub-section IV-A, CoHOG generally achieves higher place matching performance while using larger image resolutions, albeit at a considerable increase in t_{VPR}. Thus, in comparison with VPR techniques such as HOG and GIST which perform better whilst utilising a lower image resolution, the ratio for CoHOG is considerably lower. HOG achieves the highest ratio using either 32x32 px or 128x128 px, depending on the dataset. For GIST, the highest ratio is obtained on the 16x16 px resized datasets, with the exception of Gardens Point and Nordland, where this is achieved on the 32x32 px resized datasets, as seen in Fig. 6.

V. Conclusions

This paper presents an in-depth study on the effects of image resolution on the place matching performance of several well-established handcrafted VPR techniques. We confirmed that local feature descriptors are unable to operate on very small images, hence the focus of this work is mostly related to global feature descriptors. We utilise several VPR datasets to present our results, and show that the time required to perform VPR is reduced with a decrease in image resolution. Moreover, this paper performs a trade-off analysis between performance and computation, showing how utilising a lower image resolution results in a more efficient VPR process to allow efficient deployment on low-end commercial products.

Apart from the computational benefits of utilising low-resolution images, this research also has potential benefits for visual privacy. Thus, VPR could potentially be performed on images of sufficiently low-resolution that they do not compromise visual privacy. An extension of this work can investigate VPR using low-resolution images in environments where delicate visual information is present, such as faces in crowded environments and car plate numbers.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

[2] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, K. Alexis, and K. McDonald-Maier, "Are state-of-the-art visual place recognition techniques any good for aerial robotics?," *arXiv preprint arXiv:1904.07967*, 2019.

[3] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. Milford, and K. D. McDonald-Maier, "Visual place recognition for aerial robotics: Exploring accuracy-computation trade-off for local image descriptors," in *2019 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, pp. 103–108, 2019.

[4] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[5] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 2196–2203, IEEE, 2009.

[6] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in manhattan world," in *ICRA Omnidirectional Vision Workshop*, pp. 4042–4047, 2010.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.

[8] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2136–2174, 2021.

[9] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "Cohog: A light-weight, compute-efficient, and training-free visual place recognition technique for changing environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1835–1842, 2020.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features.," in *Computer Vision and Image Understanding - CVIU*, vol. 110, pp. 404–417, 01 2006.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 vol.2, 1999.

[13] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *The international Journal of robotics Research*, vol. 21, no. 8, pp. 735–758, 2002.

[14] H. Andreasson and T. Duckett, "Topological localization for mobile robots using omni-directional vision and local features," *IFAC Proceedings Volumes*, vol. 37, no. 8, pp. 36–41, 2004.

[15] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1281–1298, 2011.

[16] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[17] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3304–3311, IEEE, 2010.

[18] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*, pp. 2564–2571, Ieee, 2011.

[19] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," *arXiv preprint arXiv:1805.07703*, 2018.

[20] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4297–4304, IEEE, 2015.

[21] S. Skrede, "Nordland dataset," 2013. Available online at: *https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/*.

[22] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[23] R. Sahdev and J. K. Tsotsos, "Indoor place recognition system for localization of mobile robots," in *2016 13th Conference on computer and robot vision (CRV)*, pp. 53–60, IEEE, 2016.