Sequence-Based Filtering for Visual Route-Based Navigation

Analyzing the Benefits, Trade-Offs and Design Choices

Tomita, Mihnea Alexandru; Zaffar, M.; Ferrarini, Bruno ; Milford, Michael J.; McDonald-Maier, Klaus; Ehsan, Shoaib

DOI
10.1109/ACCESS.2022.3196389

Publication date
2022

Document Version
Final published version

Published in
IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 July 2022, accepted 30 July 2022, date of publication 4 August 2022, date of current version 10 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3196389

RESEARCH ARTICLE

Sequence-Based Filtering for Visual Route-Based Navigation: Analyzing the Benefits, Trade-Offs and Design Choices

MIHNEA-ALEXANDRU TOMIȚĂ[1], MUBARIZ ZAFFAR[2], BRUNO FERRARINI[1], (Student Member, IEEE), MICHAEL J. MILFORD[3], (Senior Member, IEEE), KLAUS D. MCDONALD-MAIER[1], (Senior Member, IEEE), AND SHOAIB EHSAN[1], (Senior Member, IEEE)

[1]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K.
[2]Cognitive Robotics Department, Delft University of Technology, 2628 Delft, The Netherlands
[3]School of Electrical Engineering and Computer Science, The Queensland University of Technology, Brisbane, QLD 4000, Australia

Corresponding author: Mihnea-Alexandru Tomiță (matomi@essex.ac.uk)

This work was supported by the U.K. Engineering and Physical Sciences Research Council under Grant EP/R02572X/1 and Grant EP/P017487/1.

ABSTRACT Visual Place Recognition (VPR) is the ability to correctly recall a previously visited place using visual information under environmental, viewpoint and appearance changes. An emerging trend in VPR is the use of sequence-based filtering methods on top of single-frame-based place matching techniques for route-based navigation. The combination leads to varying levels of potential place matching performance boosts at increased computational costs. This raises a number of interesting research questions: How does performance boost (due to sequential filtering) vary along the entire spectrum of single-frame-based matching methods? How does sequence matching length affect the performance curve? Which specific combinations provide a good trade-off between performance and computation? However, there is lack of previous work looking at these important questions and most of the sequence-based filtering work to date has been used without a systematic approach. To bridge this research gap, this paper conducts an in-depth investigation of the relationship between the performance of single-frame-based place matching techniques and the use of sequence-based filtering on top of those methods. It analyzes individual trade-offs, properties and limitations for different combinations of single-frame-based and sequential techniques. The experiments conducted in this study demonstrate the benefits of sequence-based filtering over the single-frame-based approach using various VPR techniques. We found that applying sequence-based filtering to a lightweight descriptor can enable higher VPR accuracy than state-of-the-art methods such as NetVLAD, while running in shorter time. For example, matching a sequence of 16 images, CALC descriptor outperforms NetVLAD on Campus Loop dataset while taking about 22% less time to perform VPR.

INDEX TERMS Sequence-based filtering, visual localization, visual place recognition.

I. INTRODUCTION

The goal of a visual place recognition (VPR) system is to determine if a currently observed place has been previously visited by a robot/human. Despite the efforts made by the research community, the VPR process remains perfectible. Dynamic environments and different poses of a robot's camera cause a place to change its appearance, rendering VPR a challenging task, as seen in Fig. 1.

In recent years, it has been shown that sequence-based VPR systems such as [1]–[4] and [5] can achieve good performance in changing environments. Thus, an almost parallel

The associate editor coordinating the review of this manuscript and approving it for publication was Wen Chen.

track has emerged where sequence-based techniques have been shown to outperform single-frame-based techniques. More importantly, the benefits presented by sequential information are generally extendable to most non-learning and learning-based VPR techniques albeit at varying levels and costs. Therefore, it is critical to understand the properties of sequential-based filtering, its trade-offs and how to deploy them on single-frame-based VPR techniques for designing better VPR systems.

To the best of our knowledge, there is no previous work that has examined this important problem in a systematic way (such as performance boost variations due to sequential filtering along the entire spectrum of single-frame-based VPR methods, the effects of sequence length on performance, performance-computation trade-off etc). To bridge this research gap, this paper investigates the relationship between the performance of single-frame-based, learnt and non-learnt VPR methods, and the use of sequence-based filtering on top of these methods. In particular, this paper introduces sequential information into a number of VPR techniques to improve conditional invariance and shows that sequence matching takes a poorly performing single-frame-based VPR technique and improves its performance. While sequence matching has a positive effect on VPR accuracy, it increases the time required to perform VPR. This paper examines the effects of different sequence lengths on the resulting performance boost and determines the optimal combinations between different VPR techniques and sequence lengths, taking into consideration both the performance and computational load of each system. We found that high-precision VPR systems sightly improve their performance from introducing sequential-based filtering. On the contrary, less accurate but lightweighted techniques can receive a significant boost in their VPR accuracy, whilst in some cases also keeping the matching time shorter than state-or-the-art techniques. For example, CALC outperforms NetVLAD on Campus Loop dataset using a sequence of 16 images while taking about 78% of the time to perform VPR.

In summary, our work provides the following contributions:

- The application of sequence-based filtering on top of single-frame-based methods is investigated. In particular, we analyzed the VPR performance improvement and the computational effort required to execute VPR using a sequence compared with single-frame approach.
- The trade-off between VPR accuracy and computational efficiency is examined, showing how lightweight techniques can replace state-of-the-art descriptors to perform VPR more efficiently, without any loss in accuracy.

The remainder of this paper is organised as follows: Section II presents an overview of the literature regarding VPR. Section III presents our implementation of sequential-based filtering on top of single-frame-based methods. Section IV describes the experimental setup for performing the analysis on trade-offs of sequential filtering for VPR.

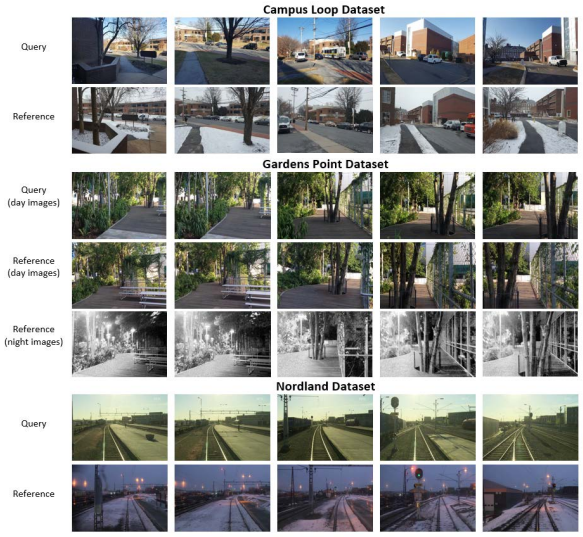

FIGURE 1. Sample sequence of images taken from each of the 4 datasets: Campus Loop, Gardens Point (day-to-day), Gardens Point (day-to-night) and Nordland (summer-to-winter).

Section V presents the detailed results and analysis. Finally, the conclusions are presented in Section VI.

II. LITERATURE REVIEW

A thorough review of existing research, current challenges and the application of Visual Place Recognition (VPR) are presented by Lowry *et. al* in [38]. Table 1 presents the highlights and limitations for some of the methods discussed in this section.

Early techniques used in the field of VPR were based on handcrafted feature descriptors [39], [40] which can be categorised into either local or global feature descriptors depending on how they extract the information from an image [38]. Local feature descriptors, such as Scale-Invariant Feature Transform (SIFT) [6] and Speeded-Up Robust Features (SURF) [7] have been used to solve VPR problem such as in [8]–[11] and [12]. FAB-MAP (Frequent Appearance Based Mapping) [41] is a VPR system that represents visual places as words and uses SURF for feature detection. The system is successfully able to deal with perceptual aliased images and can perform loop-closure detection. CAT-SLAM [42], extends the work of FAB-MAP by including odometry information. Center Surround Extremas (CenSurE) [43] performs real-time detection and matching of image features and has been employed by FrameSLAM in [44]. The Bag-of-Words model (BoW) [13] has been used for VPR tasks such as in [45]. Another important handcrafted technique is the Vector of Locally Aggregated Descriptors (VLAD) [14]. Both [13] and [14] are used to partition the feature space in a fixed number of visual words, that enables more efficient image matching. A popular whole-image descriptor is Gist [15], [16] which has been used in [21], [22] and [18] for image matching. BRIEF has been paired with Gist by the authors of [18]. Histogram-of-Oriented-Gradients

TABLE 1. Summary of the highlights and limitations of a selection of VPR methods available in the literature.

Method	Description	Highlights	Limitations
SIFT [6] SURF [7]	- Local image feature descriptor. - Handcrafted techniques used for VPR in [8], [9], [10], [11] and [12].	- Locality. - Complete pipeline from key-point detection to description.	- The whole image descriptor is very large. - Image matching can be inefficient if the descriptor is not used with a feature aggregator such as BoW [13] or VLAD [14].
GIST [15] [16] BRIEF [18] HOG [19] [20]	- Global image descriptors. - GIST was used in VPR applications in [21], [22], [18], BRIEF in [18] and HOG in [23] and [24].	- Compact. - Fast to perform image matching.	- sensitive to viewpoint changes [17]
HybridNet [25] AMOSNet [25]	- CNN based on AlexNet [26] trained on scene themed datasets such as SPED [25]. - Designed for VPR.	- Tolerant to environmental changes such as seasonal and illumination variations.	- Poor performance on viewpoint changes.
NetVLAD [27]	- CNN-based descirptor including a VLAD-like layer, designed for VPR.	- Trained in weakly supervised manner, using triplet loss function to achieve viewpoint tolerances.	- Relatively low performance with environmental changes.
CALC [28]	- CNN trained in an unsupervised manner to reproduce HOG from distorted place images. - Designed for VPR.	- More efficient than many other CNN-based methods.	- Low VPR accuracy compared to other CNN-based VPR techniques.
SeqSLAM [1]	- Sequence-based VPR method. - Handcrafted technique designed for VPR.	- Very robust to environmental changes.	- Very sensitive to viewpoint changes. - Does not perform well with variable velocities.
SMART [29]	- Sequence-based VPR method that incorporates odometry information to improve VPR. - Handcrafted technique designed for VPR.	- It is invariant to illumination variation and vehicle speed. - Better performance than SeqSLAM.	- The system does not perform well on datasets that contain a huge amount of similar frames (aliased images).
VPR method for aerial robots [30]	- Sequence-based VPR system.	- It does not require the query image sequences to have the same order as the database images.	- The performance of the system has not been determined at different altitudes.
Visual Localisation using Network Flow [31]	- Sequence-based VPR system.	- It is able to deal with substantial seasonal changes. - It is able to operate with variable velocities.	- At low recall rates, the proposed technique has low accuracy.
Fast, compact and highly scalable VPR [32]	- Sequence-based VPR pipeline.	- Low overall storage footprint, extremely fast retrieval and sub-linear storage growth.	- Clustering is imbalanced when no re-scaling is performed, leading to poor performance.
Feature Co-occurrence Maps [33]	- Appearance-based localisation across multiple times of day.	- It is successfully able to perform VPR under illumination variation at high precision/recall.	- Sequence matching struggles with dynamic objects.
ConvSequential-SLAM [5]	- Handcrafted and sequence-based VPR technique.	- It achieves state-of-the-art performance on viewpoint and appearance variant datasets.	- The proposed technique is not able to deal with dynamic objects and confusing features.
DeepSeqSLAM [3]	- A sequence-based CNN+RNN system.	- It is successfully able to outperform classical sequence-based methods in accuracy, run-time and computational requirements.	- The CNN component of the system is not able to generalize to drastic visual changes.
HMM Sequence Matching [34]	- Sequence-based VPR technique.	- It is able to deal with non-linear changes in velocity. - It outperforms SeqSLAM.	- The system can only deal with small variations in viewpoint.
STA-VPR [35]	- Sequence-based wrapper method.	- It is successfully able to deal with variable velocities and it is robust to appearance and viewpoint variations.	- The method is outperformed by SSM-VPR [36] [37] under drastic viewpoint variations.

(HOG) [19], [20] is another whole-image descriptor used by the authors of [23]. Zaffar *et al.* [24] present a handcrafted VPR technique which employs HOG feature descriptors to achieve state-of-the-art place matching performance in changing conditions. The proposed approach has zero training requirements and low encoding times, hence it is a great alternative to more resource-intensive VPR techniques, especially for deployment on resource constrained robotic platforms. The use of complementary of VPR techniques is an emerging approach to address VPR. The work presented in [46] examines the strengths and weakness of various VPR approaches and optimal combinations of methods are proposed for different environmental conditions. SwitchHit [47] relies on complementary to propose a switching system to select the optimal VPR algorithm in dynamic environments.

Convolutional Neural Networks (CNNs) have been widely explored by researchers (such as in [48] and [49]) in VPR. Chen *et al.* [50] used the spatial filter of SeqSLAM together

with all the layers of the Overfeat Network [51]. The authors of [25] created two neural-network based VPR techniques. The first architecture, entitled HybridNet, used weights learnt from the top 5 convolutional layers of CaffeNet [52], while the second architecture, AMOSNet, was trained from scratch on the SPED dataset. The authors of NetVLAD [27] presented a new Vector-of-Locally-Aggregated-Descriptors (VLAD) layer that can be incorporated in any neural network architecture, drastically enhancing the performance in VPR related scenarios. Merrill *et al.* [28] showed that convolutional auto-encoders are suitable for VPR tasks. The resulting CNN, CALC, is lightweight as well as robust to variations in both illumination and viewpoint. However, CALC has low accuracy when compared to other CNN-based VPR techniques. Cross-Region-Bow [53] achieves viewpoint tolerance by building an image representation from a pre-trained CNN. First, it searches for local maxima in a pre-trained CNN's feature map to identify regions of interest (ROI). Then, the features underlying the selected ROIs are pooled to form an image descriptor using BoW. RegionVLAD [54] is based on the same approach as Cross-Region-BoW but employs VLAD for feature pooling. [55] and [56] present computationally efficient and compact binary neural networks (BNN) for VPR achieving comparable performance in changing environments with full-precision systems such as HybridNet. However, BNNs require dedicated hardware or an inference engine that enables an efficient computation of bitwise operations. Bio-inspired algorithms are considered as well to address VPR efficiently. Arcanjo *et al.* [57] proposed a lightweight network inspired by Drosophila neural system consisting in a pre-processing stage to compute a compact binary image representation, followed by a classifier to predict the current location of a robot.

SeqSLAM [1] performs visual place recognition in changing environments by comparing sequences of camera frames instead of the conventional single-image approach, in order to decide whether the place has been previously visited. To achieve sequence matching, SeqSLAM uses a set of pre-defined constant velocity search lines through the difference matrix in order to break the map into multiple places. SMART [29] extended SeqSLAM by incorporating the odometry into its calculations. The authors of [30] proposed a new sequence-based VPR system for aerial robots. This method uses Bayes estimation to perform sequential image matching and it does not require that the sequence of query images to be organised in the same order as the stored map. In [32], a fast and compact VPR pipeline is presented where sequence matching is used to resolve the collisions in the hash space. Johns *et al.* [33] show a new method for appearance-based localisation, namely Feature Co-occurrence Maps. The performance of this technique does not degrade during severe changes in illumination, thus place matching is performed at high precision/recall. Co-occurrence Maps outperforms both FAB-MAP [41] and SeqSLAM [1]. The authors of [31] propose a sequence-based VPR system with robust localisation that can deal with substantial seasonal changes. More recently, in [5], the authors have presented a sequence-based VPR system based on HOG descriptors that is able to perform place matching in challenging conditions, using adaptive sequence-based matching to tackle VPR in dynamic environments. DeepSeqSLAM [3] is a trainable CNN+RNN system that is successfully able to complete VPR related tasks in challenging environments. The authors of [34] propose a VPR algorithm that matches sequences of query and reference frames. A matrix of low-resolution, contrast-enhanced image similarity values are computed in order to perform sequence matching and a Hidden Markov Model (HMM) framework is used to find the best sequence alignment. STA-VPR [35] is a sequence-based VPR technique that uses an adaptive dynamic time warping (DTW) algorithm in order to improve its robustness to changes in appearance and viewpoint. Furthermore, to achieve image sequence matching based on temporal alignment, a local matching DTW (LM-DTW) algorithm is used, thus achieving a linear time complexity. Both [34] and [35] are suitable to deal with non-linear changes in velocity, whereas [1] does not perform well with variable velocities.

Algorithm 1 Query and Reference Descriptor Comparison

Given: Query Descriptor (Q_F)
Given: Map of Reference Descriptors (R_M)
INITIALISE (array of 0s): score_array[length(R_M)]
iterator = 0
for R_F ***in*** R_M **do**
    score = Cosine_Similarity(Q_F, R_F)
    score_array[iterator] = score
    iterator = iterator + 1
Best_Match = Max(score_array)

III. METHODOLOGY

This section presents the approach taken for evaluating the boost in performance resulted from introducing sequential-based filtering on top of single-frame-based techniques. To enable the comparison of different VPR descriptors, the sequence filtering schema presented in [5] has been employed, as it is agnostic to the underlying single-frame technique. This approach combines the outcome of single-frame matching operations into a scalar symbolizing the similarity between sequences of images representing the places to match. The below sub-sections provide details on sequential-based filtering and the evaluation criteria used to assess the impact of sequential filtering on single-frame-based VPR techniques.

A. SINGLE-BASED IMAGE MATCHING

For any given query image (e.g. a frame taken from a robot's camera), the main goal of a VPR technique is to retrieve the most representative reference image (the matching place) from the database. This is done by comparing each query image with all the stored database images in such a way that each time a query and reference image are matched together,

a similarity score is computed. For any given query image, the reference image with the highest score is chosen as the best match.

The feature descriptor computed by a VPR technique for a query image Q is denoted as Q_F, for a reference image R as R_F whilst the list containing the reference descriptors for the entire map as R_M. The similarity between two image descriptors (Q_F and R_F) is determined using the cosine [53]:

$$s = \frac{Q_F \dot{R}_F}{||Q_F|| * ||R_F||} \tag{1}$$

The single frame-matching schema requires that Q_F is compared with every R_F from R_M. Thus, for any N images in a dataset, a set of similarity scores S is created as follows:

$$S = \{s_1, s_2, s_3, \ldots, s_N\} \tag{2}$$

where $s \in \mathbb{R}$ and s in range [0,1]. Higher the score, higher the similarity between two image descriptors.

For each query image Q, a new set of similarity scores S is created containing the values for that particular frame. Once the similarity coefficients have been computed, the reference image with the highest value ($s \in S$) is regarded as the matching place for Q_F.

Algorithm 1 presents the entire matching process for a query image descriptor Q_F and the map, R_M. The matching score (calculated as in equation (1)) of each query-reference pair is stored in a 1D array entitled *score_array*. Once a similarity score has been generated for every R_F (from R_M), the maximum value from the *score_array* is retrieved, and thus, the most representative reference image for Q_F is selected as the best match.

B. SEQUENTIAL-BASED FILTERING

In contrast to the single-image matching process previously mentioned, sequential-based filtering allows a VPR technique to match sequences of query and reference frames. The most important steps for introducing sequential-based filtering on top of single-frame-based VPR techniques are presented below:

1) CREATING THE IMAGE SEQUENCE

For any given query image q_i, the sequence of **K** consecutive images is built as follows:

$$q_i \; q_{i+1} \; q_{i+2} \; \cdots \; q_K \tag{3}$$

where q_i is the query image for which the sequence is built, q_K is the last query image that is part of the given sequence, and **K** is the total number of images that forms each sequence.

Similarly to (3), the reference images are organised in sequences (formed with an offset of 1 image) as presented in equation (4):

$$\begin{matrix} r_1 & r_2 & r_3 & \cdots & r_K \\ r_2 & r_3 & r_4 & \cdots & r_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{N-K+1} & r_{N-K+2} & r_{N-K+3} & \cdots & r_N \end{matrix} \tag{4}$$

Algorithm 2 Creating Query and Reference Sequences

Given: Total Number of Query Images
Given: Total Number of Reference Image
K = image sequence length
for *i* ***in range*** *(total_Query_Images* $-$ ***K*** $+ 1$*)* **do**
  ref_matching_scores = []
  for *j* ***in range*** *(total_Ref_Images* $-$ ***K*** $+ 1$*)* **do**
    score = *perform_VPR*(Q_F, R_F, *i*, *j*)
    ADD score to ref_matching_scores
  Best Match = Max (ref_matching_scores)

Algorithm 3 The *perform_VPR* Function Is Presented Here

Given: List of Query Descriptors (Q_F)
Given: List of Reference Descriptors (R_F)
Given: Query Image Number (*i_query*)
Given: Reference Image Number (*j_ref*)
K = image sequence length sequential_score = 0
i = i_query
j = j_ref
while *i* $<$ *i_query* $+$ ***K*** *and j* $<$ *j_ref* $+$ ***K*** **do**
  score = Cosine_Similarity(Q_F[i], R_F[j])
  sequential_score = sequential_score + score
  i = i + 1
  j = j + 1
sequential_score = sequential_score / **K**

The application of equation (4) results in $N - \boldsymbol{K} + 1$ image sequences, where N is the total number of images in the dataset and **K** is the sequence length. Using higher sequence lengths will lead to less images to be searched for, as no new image sequences of length **K** can be created when we approach the end of the dataset. For this reason, the number of sequences created depends solely on the value of the selected sequence length ($2 \leq \mathbf{K} \leq N$) as shown below:

$$No.\ of\ Seq\ Created = N - K + 1 \tag{5}$$

Once the query and reference sequences are created, the sequence matching is performed.

2) SEQUENCE MATCHING

All query and reference features are initially computed and stored in two separate 1D lists: Q_F and R_F. *perform_VPR* in Algorithm 2 has two main functions, more specifically creating the query and reference image sequences of constant length **K** from Q_F and R_F (presented in sub-section III-B1) and image sequence matching.

The *perform_VPR* function firstly takes the indices (*i* for query images and *j* for reference images) from Algorithm 2 in order to determine for which query and reference image the sequences will be created. Starting from the *i-th* image, the *perform_VPR* function creates sequences by adding consecutive images until the required sequence length **K** has been obtained. The same process is repeated for every reference

image, starting with the *j-th* image. This process is presented in Algorithm 3, which represents the *perform_VPR* function.

For every given query image sequence previously created, *perform_VPR* searches for the most representative reference image sequence. Algorithm 3 presents the process of matching a sequence of query and reference images, generating **K** similarity values (*score*) for each query-reference pair that are part of the matched sequences. The similarity or matching score of any query-reference image sequences (*sequential_score*) is calculated as the arithmetic mean of the matching scores of the pairs within these sequences. Thus, the matching score for a sequence of images of length **K** is computed as:

$$s' = \frac{\sum_{i=1}^{K} s_i}{K} \tag{6}$$

where s_i represents the matching score for each query-reference pair with index i. The matching score s' has values in range [0, 1], with a higher score denoting a better similarity between two sequences of query and reference frames. Thus, for each query image sequence, the reference sequence with the highest score is selected as the most representative match. This can be seen in Algorithm 2, where for any given query image sequence, the matching scores of all reference image sequences are stored in a list, namely *ref_matching_scores*. The maximum score from this list is taken as the best match for that given query image sequence.

When analysing a query image q_i, we take into account the sequential information provided from using consecutive images, thus the next $\boldsymbol{K} - 1$ images are also analysed as part of q_i's image sequence. For this reason, the first reference image that is part of the sequence with the highest score is retrieved as being the best match for its corresponding query image.

IV. EXPERIMENTAL SETUP

This section discusses the performance metrics employed, the VPR techniques utilised to generate our results and the sequential datasets used in this work.

A. EMPLOYED PERFORMANCE METRICS

Area-under-the-Precision-Recall-Curve (AUC) is widely used in VPR research for evaluation purposes [17] due to the fact that it performs well on unbalanced data, which is also the case for VPR applications. Thus, it is also employed in this work utilizing (7) and (8):

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{7}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{8}$$

Another important metric utilised in our work is the accuracy [5] with the following definition:

$$A = \frac{No.\ of\ Correctly\ Matched\ Query\ Images}{Total\ No.\ of\ Query\ Images\ in\ Database} \tag{9}$$

The authors of [28], [54], [58] and [59] determined that the feature encoding time (t_e) of a VPR system to be an important performance indicator. In [24], the authors evaluated a system's performance using Performance-per-Compute-Unit (PCU). This is defined by combining precision at 100% recall (P_{R100}) with t_e as in equation (10):

$$PCU = P_{R100} \times \log\left(\frac{t_{e\_max}}{t_e} + 9\right) \tag{10}$$

In this equation, the maximum feature encoding time (t_{e_max}) is used to represent the most resource intensive VPR technique, while t_e represents the feature encoding times for each of the remaining techniques (where $t_e < t_{e\_max}$). It is worth mentioning that without the scalar 9 in equation (10), the VPR technique with $t_e = t_{e\_max}$ will always result in a PCU of 0. Techniques with higher precision and lower feature encoding time generally lie towards the higher spectrum of PCU, while compute-intensive and less precise techniques converge towards lower PCU values. Thereby, this addition provides a more interpretable range. Because PCU is a relative performance metric, it serves us value in this study.

B. UTILISED VPR TECHNIQUES

In this work, sequence-based filtering is introduced into a number of state-of-the-art VPR techniques, namely HOG [20], CALC [28], AMOSNet [25], HybridNet [25] and NetVLAD [27]. Single-frame-based implementation of Zaffar *et. al* [17] is used for all 5 aforementioned VPR techniques. In Section V, comparative results based on the above-mentioned performance metrics for these VPR techniques are presented along with discussion of the benefits and trade-offs of sequence-based filtering.

C. UTILISED SEQUENTIAL DATASETS

For this study, four sequential VPR datasets are used. The first dataset is Campus Loop dataset [28], which contains 100 query and 100 reference images. It poses challenges to any VPR system due to the high amount of viewpoint variation, seasonal variation and also the presence of statically-occluded frames. The second and third datasets are part of Gardens Point dataset [26] which contains both day and night images, that are divided as follows: 200 query images (*day left*) and 400 reference images (equally split into day images (*day right*) and night images (*night right*)). Nordland dataset [60] is the fourth dataset used which captures the drastic visual changes that seasonal variation can have on a place (spring, summer, autumn and winter). Since the most notable differences between seasons are seen during the summer and winter seasons, each VPR technique is tested here on the summer-to-winter traverses of the Nordland dataset. Fig. 1 shows sample images taken from each dataset.

V. RESULTS AND ANALYSIS

In this section, we focus our attention on understanding how and when to use sequence-based localisation/place

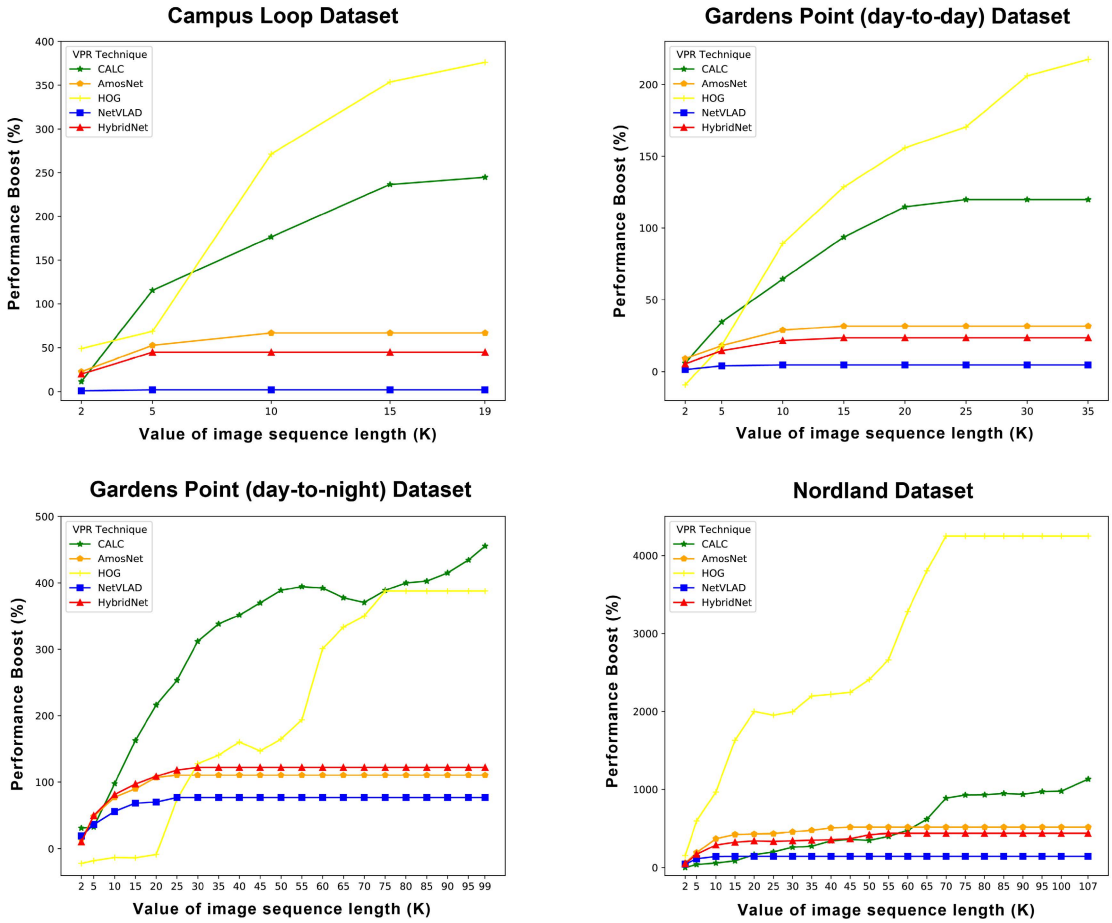

FIGURE 2. The performance boost (%) of sequence matching performance in comparison to the single-frame-matching performance of all VPR techniques on the datasets mentioned in sub-section IV-C.

matching, its strengths/downsides and the most appropriate performance metrics that can be used in order to test the efficacy of such VPR systems. We present the results for sequence-based filtering when used on top of the VPR techniques mentioned in sub-section IV-B. We also present the computational effects of sequential-based filtering and discuss the benefits and trade-offs. For all experiments presented below, we have used a PC equipped with an Intel Core i7-4790k CPU.

A. PLACE MATCHING PERFORMANCE

Fig. 2 presents the performance boost provided by sequence matching for several sequences lengths. The upper limit of **K** shown in Fig. 2 is determined by reaching 100% accuracy by every method. Those **K** values are summarized in Table 2 for every VPR technique and dataset. The performance boost in Fig. 2 is calculated as the percentage increase between the accuracy of the sequence-based and the single-image version of the same VPR technique. It is evident from Fig. 2 that the addition of sequential filtering to a given single-frame-based VPR technique mostly improves the overall place matching performance of that technique. This suggests that by increasing the sequence length of a VPR technique, we will achieve better place matching performance. HOG achieves the highest performance boost on all datasets except Gardens Point day-to-night. VPR techniques such as AMOSNet and HybridNet have a substantial increase in performance using a considerable shorter sequence length (**K**) than simpler VPR techniques, such as CALC or HOG, on Gardens Point day-to-night and Nordland. The reason behind this is that CNN-based VPR techniques such as AMOSNet and HybridNet are designed and trained to deal with drastic changes in the environment, while simpler techniques such as HOG are only able to deal with moderate viewpoint and illumination changes. We further discuss this topic in sub-section V-B. However, VPR techniques which already achieve close-to-ideal matching performance, such as NetVLAD, do not benefit much from using an increased sequence length on certain datasets, such as on the Campus Loop and Gardens Point day-to-day dataset, where the performance boost of the system is negligible. This is mainly because CNNs such as NetVLAD, are successfully able to handle the viewpoint, seasonal and illumination variations that can be found in these datasets, without requiring an increased sequence length. This observation is important as using sequences instead of single images has computational drawbacks and should be

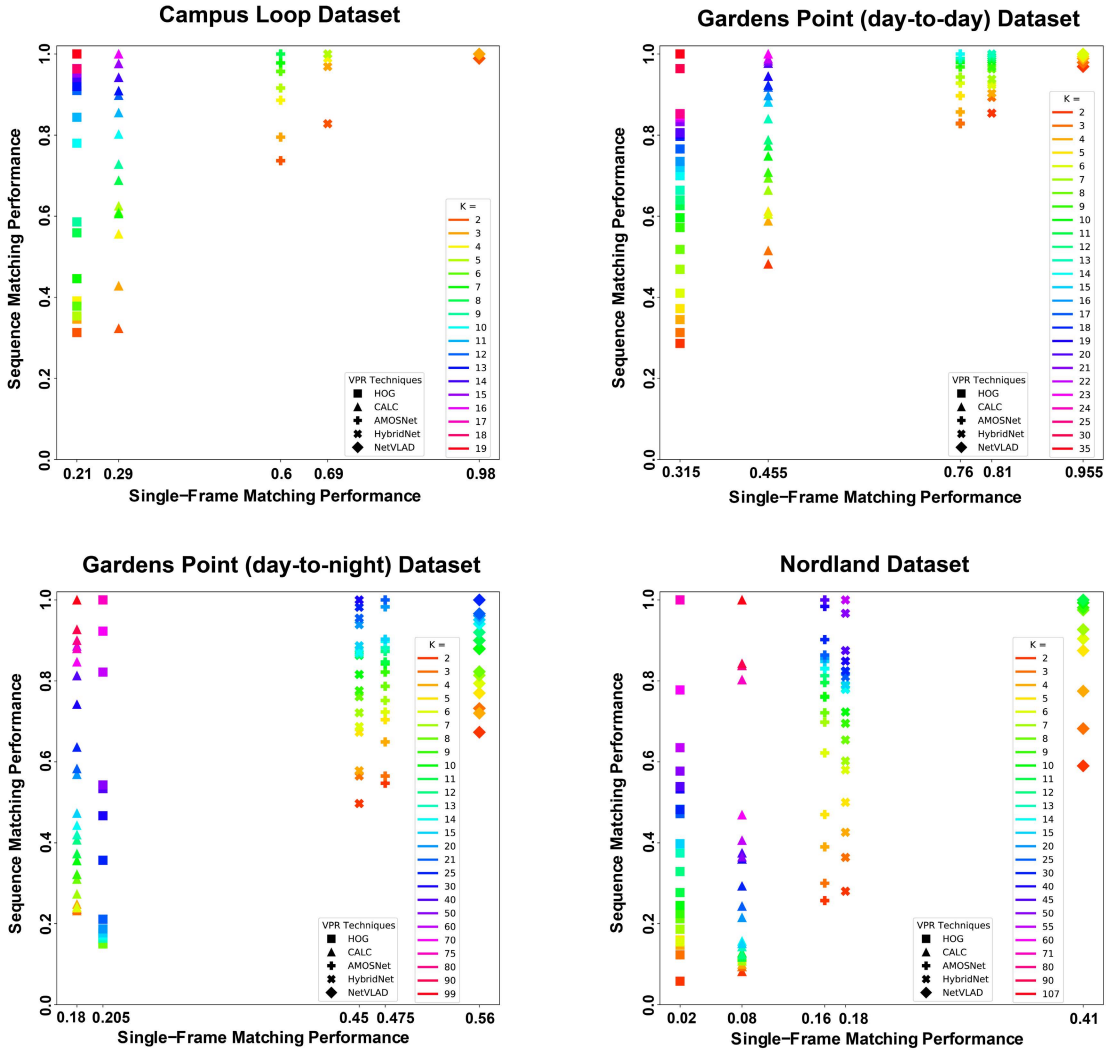

FIGURE 3. The single-frame matching performance compared to the sequence matching performance for all 5 VPR techniques on all 4 datasets.

avoided where unnecessary. We expand on this further in sub-section V-C1 and V-D.

B. PERFORMANCE-BOOST VARIATIONS

Fig. 3 shows the performance boost provided by sequence matching (y-axis) compared to the single-frame accuracy, which is reported on the x-axis for each of the considered VPR techniques. The values are reported only up to the sequence length that enables a perfect score. For example, HOG achieves 100% accuracy on Campus Loop for a sequence length of 19 images.

A common observation in existing literature has been that sequential-filtering mostly helps with introducing conditional-invariance [1], however, the results obtained on Gardens Point day-to-day shown in Fig. 3 demonstrate that it also greatly helps in viewpoint-variant, conditionally-invariant scenarios. The performance boost of each VPR technique is directly linked to the intensity of conditional variations (and their effects on the scene appearance) in the dataset. The benefits of sequential-filtering are clearly enjoyed extensively by most techniques on datasets (Campus Loop and Gardens Point day-to-day) with less conditional changes than datasets (Nordland and Gardens Point day-to-night) with extreme conditional changes. Fig. 6 shows a sequence of correctly matched query and reference images taken from each of the 4 datasets.

In contrast to the observations made above, the performance improvement of HOG (refer to Fig. 2) is inconsistent on the Gardens Point day-to-night dataset (for sequence lengths of $2 \leq \mathbf{K} \leq 20$), where the single-frame performance of this technique achieves similar or better place matching performance compared to that of the sequence matching performance. The presence of extreme viewpoint variation, illumination variation and also the presence of statically-occluded frames in the Gardens Point day-to-night dataset may affect the performance of this technique. Similarly, the improvement in the performance gained by using sequential-based matching for CALC is more limited (thus

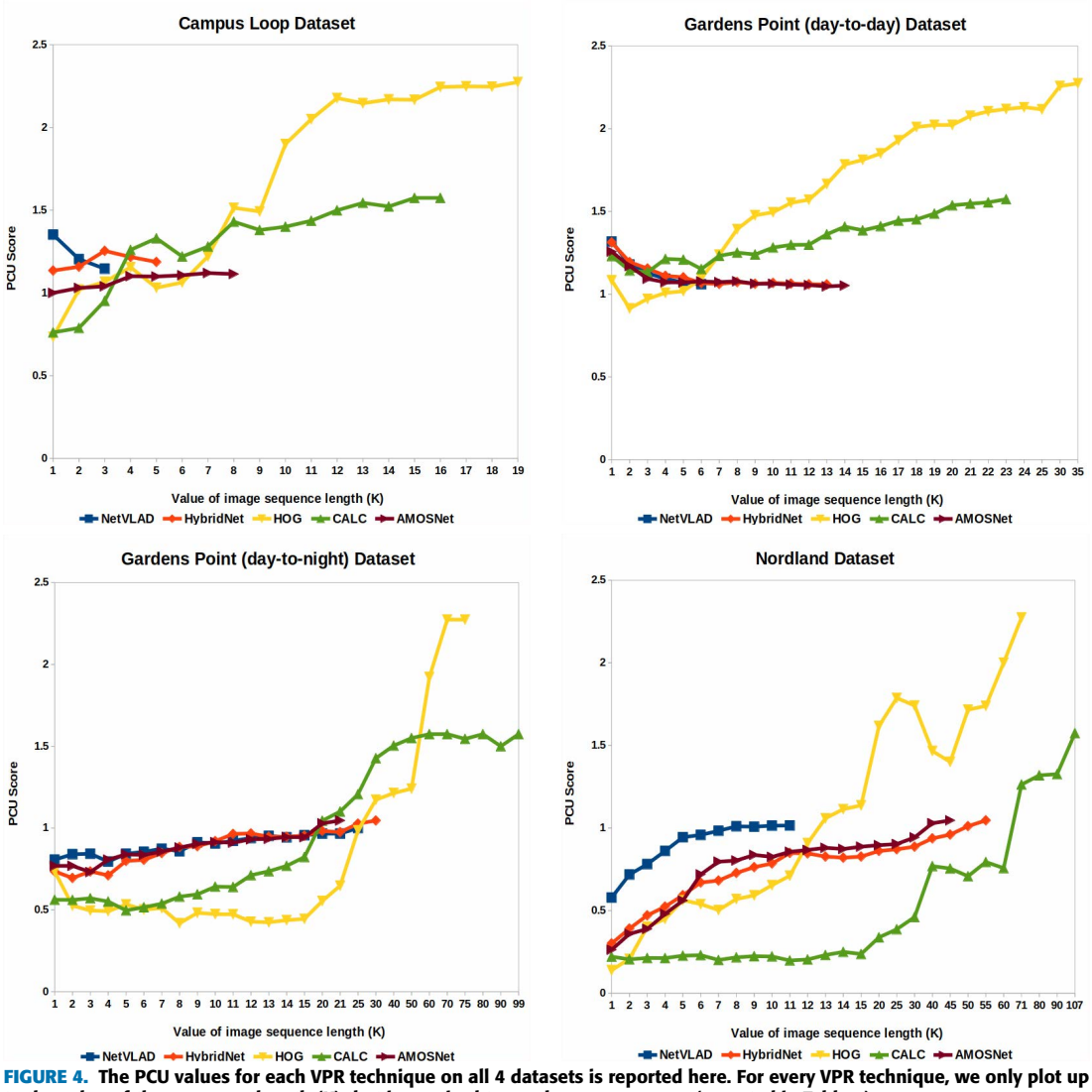

FIGURE 4. The PCU values for each VPR technique on all 4 datasets is reported here. For every VPR technique, we only plot up to the value of the sequence length (K) that is required to reach 100% accuracy (reported in Table 2).

requiring a longer sequence length **K** to reach maximum accuracy) when compared to other techniques on the Nordland dataset due to the presence of viewpoint and seasonal variation, as seen in Fig. 3. On this dataset, even with the addition of sequential-based matching, both HOG and CALC achieve lower results than more complex VPR techniques such as NetVLAD. These results are primarily due to the nature of the dataset, which contains a large number of confusing features, primarily coming from trees and vegetation. On the other hand, the night images from Gardens Point contain a lot of noise (pepper noise) which drastically decrease the place matching performance of light-weight systems such as HOG. We show in Fig. 7 some sequences of incorrectly matched query and reference images taken from both Gardens Point day-to-night and Nordland datasets. In such scenarios, evidently it is better for a system to switch to more sophisticated and invariant techniques, such as NetVLAD and HybridNet, even at the expense of higher computational needs.

In summary, some example cases where using a higher sequence length for a trivial VPR technique (such as HOG) is beneficial are laterally viewpoint variant and seasonally variant (but under similar illumination) scenes, e.g. driving a car in a different lane on a previously visited road in a different season. The increasing trend in performance of the HOG technique can be clearly seen in both Fig. 2 and Fig. 3, for the Campus Loop and Gardens Point day-to-day datasets. However, for platforms that can have 3D or 6-DOF viewpoint changes, e.g. drones, UAVs etc, deep-learning-based techniques should be used instead of trivial techniques with high sequence length, which is also the case for highly illumination/conditionally variant scenes such as those found in the Gardens Point (day-to-night) and Nordland datasets. Our data supports the fact that deep-learning-based VPR techniques are better equipped to deal with these variations, and that they should be used in these scenarios instead of more simple VPR systems. Thus, we propose that having this prior knowledge can lead a system based on an

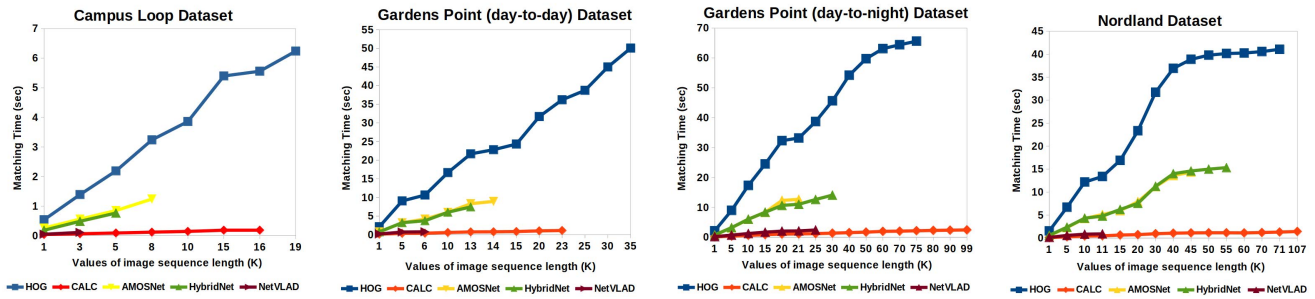

FIGURE 5. The matching time in seconds of each VPR technique on all 4 datasets is presented here. For every VPR technique, we only plot up to the value of the sequence length (K) that is required to reach 100% accuracy (reported in Table 2).

TABLE 2. The sequence length (K) required for each VPR technique to reach maximum place matching performance (100% accuracy) on each of the 4 datasets.

Dataset	VPR Technique				
	NetVLAD	HOG	CALC	AMOSNet	HybridNet
Campus Loop	3	19	16	8	5
Gardens Point (day-to-day)	6	35	23	14	13
Gardens Point (day-to-night)	25	75	99	21	30
Nordland	11	71	107	45	55

ensemble of sequentially-filtered VPR techniques, which are switched accordingly dependent upon the environmental variation cues. This criteria will ensure that the most appropriate VPR technique is selected in each scenario, thus increasing the place matching performance, possibly at much lower computational costs as discussed in sub-section V-C1.

C. BENEFITS AND TRADE-OFFS OF SEQUENTIAL FILTERING

This sub-section presents the benefits and trade-offs of sequential filtering while also answering key questions.

1) COMPUTATIONAL EFFECTS OF SEQUENTIAL-FILTERING

Due to the fact that we are matching sequences of images instead of the traditional single-frame approach, the feature encoding time for each VPR technique will be increased by **K** folds. Table 3 shows the feature encoding time of the 5 VPR techniques used in this work without sequential filtering and Fig. 5 presents the matching time of each technique. Because neural network-based VPR techniques, such as HybridNet, AMOSNet and NetVLAD already have increased feature encoding times, the addition of sequential filtering will lead to a drastic increase in processing time. Fig. 4 shows the Performance-Per-Compute-Unit (PCU) of each VPR technique and the computational effects of using multiple sequence lengths. Thus, in both Fig. 4 and 5, for each VPR technique, we only plot up to the sequence length values (**K**) that are required to achieve 100% accuracy (see Table 2). It is important to note that a significant increase in the PCU curves occurs when there is a notable increase in precision compared to the increase in encoding time. HOG achieves high PCU values due to both its low encoding times and high increase in precision when adding sequential filtering.

Apart from the computational downsides mentioned above, the latency in getting a match as it need to build up sequence has to be considered. Furthermore, shifting between two different routes that have not been traversed in that order in the map (switching latency) as well as the difficulties with variable velocities (solved partially with more sophisticated search or using odometry information) can lead to further computational constraints. This is especially important for resource constrained platforms as it may restrict its applicability in real world scenarios, due to the high amount of visual information that has to be processed.

2) SEQUENCE-BASED FILTERING VS. SINGLE-IMAGE-BASED VPR

The data shows that for a VPR system that has poor performance on a dataset, the addition of sequence-based filtering may greatly improve its performance. Using a longer sequence length will have a higher impact in place matching performance. This is the case for HOG and CALC, which greatly benefit from the addition of sequence-based filtering. On the other hand, the single-image version of NetVLAD already achieves almost perfect results on both Campus Loop and Gardens Point (day-to-day) datasets and thus, the increased computational effects of sequential filtering for just a small gain in place matching performance may not evidently be desirable, as shown in Fig. 4. Empirically, increase of sequence length does not cause any reduction in the place matching performance but mostly yields better performance and therefore, given computational power, it may be desirable to use sequence-based techniques instead of single-image-based techniques.

3) PERFORMANCE BENEFITS BASED ON SEQUENTIAL FILTERING

As shown in Fig. 2 and Fig. 3, an increased sequence length for a given VPR technique will lead to higher performance on most datasets tested. However, different VPR techniques will

TABLE 3. Feature encoding times of different VPR techniques.

VPR Technique	Feature Encoding Time (sec)
AMOSNet	0.36
CALC	0.027
HOG	0.0043
HybridNet	0.36
NetVLAD	0.77

require different sequence lengths (see Table 2) depending on the performance of the system on a given dataset. When using sequence-based filtering, the boost in performance can be attributed to several reasons. Primarily, using an increased sequence length increases the chances of finding the best reference image for any given query image which also translates to reduced perceptual aliasing. The increased sequence length also improves the conditional-invariance of a VPR technique as shown by our results.

4) LIGHT-WEIGHT VS. DEEP-LEARNING-BASED VPR TECHNIQUES BLENDED WITH SEQUENTIAL-BASED FILTERING

It is evident that it is indeed possible to use a much simpler, light-weight VPR technique, paired with sequential filtering in order to match or even outperform the effectiveness of deep-learning-based VPR techniques on certain datasets. We have shown that the performance of a simpler VPR technique, such as HOG, can be drastically increased when using sequence-based filtering with a longer sequence length. The same can be said about CALC, which achieves good results when paired with sequential filtering. Moreover, both VPR techniques have a low feature encoding time, thus greatly benefiting from a PCU standpoint. Using the best VPR techniques (simpler systems with longer sequence lengths or deep-learning-based systems with smaller sequence length) for the right dataset will result in an overall better place matching performance, as discussed in sub-section V-B.

D. COMPUTATIONAL BUDGET

In a real-word scenario where robotic platforms are computationally restrained, it is imperative to achieve the highest VPR performance given computational constraints. In this sense, we show a performance comparison between the best performing single-frame-based VPR technique and the sequence length obtainable by each VPR technique in a given time frame. By adding together the encoding time (t_e) with the matching time (t_m), we obtain the *VPR time* for any technique as follows:

$$t_{VPR} = t_e + t_m \tag{11}$$

Using equation (11) allows us to make a fair comparison between the performance of each VPR technique and the effects that sequence length (**K**) has on t_{VPR}. For this reason, t_{VPR} is used as a criterion that helps us determine whether the best performing single-frame-based VPR technique (NetVLAD - refer to Fig. 3) can be outperformed by a sequence-matching filtering implementation of other VPR techniques presented in this work. As such, given the t_{VPR} of NetVLAD as computational budget, we present in Table 4 the maximum sequence length (**K**) obtainable by each VPR technique in regards to the given time. In case where a VPR technique reaches 100% accuracy before the computational budget is expended, the respective sequence length (**K**) is reported instead. Apart from the accuracy of a VPR system, we also present the AUC values and the precision at 100% recall (P_{R100}) for that particular sequence length (**K**).

HOG has the lowest encoding time t_e of all VPR techniques presented. However, due to its increased matching time t_m, it is unable to achieve a sequence length of $\mathbf{K} > 1$ in less t_{VPR} than NetVLAD. On the other hand, CALC has an overall low t_{VPR}, thus being able to compute a longer sequence length than every other technique. We show in Table 4 that, on Campus Loop dataset, CALC is able to achieve better performance than NetVLAD, in less t_{VPR}. However, due to the low single-frame matching performance of CALC on datasets such as Gardens Point (day-to-night) and Nordland (as shown in Fig. 3), a much longer sequence length **K** than the one obtained in the given t_{VPR} would have been required

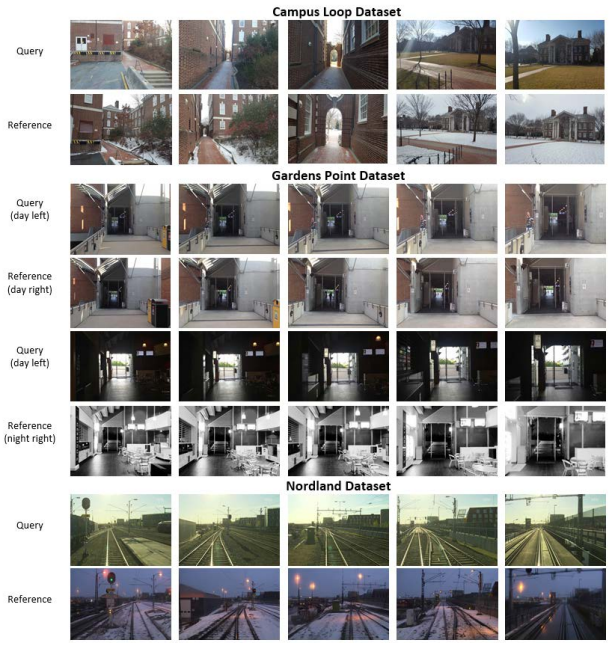

FIGURE 6. Some correctly matched sequences of query and reference images taken from each of the 4 datasets used.

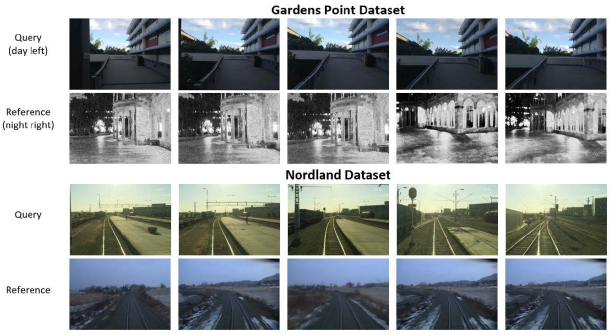

FIGURE 7. Some incorrectly matched sequences of query and reference images taken from Gardens Point day-to-night and Nordland datasets.

TABLE 4. A comparison between the best performing single-frame-based VPR technique (NetVLAD) and the maximum sequence length that can be reached by the sequence-based implementation of the remaining VPR techniques within the given computational budget (refer to sub-section V-D). In the case where a VPR technique reaches 100% accuracy before the computational budget is expended, the performance for the obtained sequence length is presented instead. The values presented in bold represent the VPR technique that has the highest accuracy and the technique that has the lowest t_{VPR} (refer to equation (11)). The comparison is performed on all 4 datasets.

	Campus Loop Dataset				
VPR Technique	NetVLAD	HOG	CALC	AMOSNet	HybridNet
K	1	1	16	1	1
t_e (sec)	0.77	0.0043	0.432	0.36	0.36
t_m (sec)	0.049	0.544	0.21	0.186	0.19
t_{VPR} (sec)	0.819	**0.544**	0.642	0.546	0.55
Accuracy	0.98	0.21	**1**	0.6	0. 69
AUC	0.998	0.301	0.999	0.872	0.889
P_{R100}	0.98	0.214	1	0.625	0.704
	Gardens Point (day-to-day) Dataset				
VPR Technique	NetVLAD	HOG	CALC	AMOSNet	HybridNet
K	1	1	10	1	1
t_e (sec)	0.77	0.0043	0.27	0.36	0.36
t_m (sec)	0.199	2.18	0.622	0.773	0.78
t_{VPR} (sec)	0.969	2.184	**0.892**	1.133	1.14
Accuracy	**0.955**	0.315	0.75	0.76	0.81
AUC	0.959	0.431	0.899	0.907	0.933
P_{R100}	0.955	0.316	0.748	0.779	0.814
	Gardens Point (day-to-night) Dataset				
VPR Technique	NetVLAD	HOG	CALC	AMOSNet	HybridNet
K	1	1	10	1	1
t_e (sec)	0.77	0.0043	0.27	0.36	0.36
t_m (sec)	0.223	2.13	0.632	0.768	0.779
t_{VPR} (sec)	0.993	2.134	**0.902**	1.128	1.139
Accuracy	**0.565**	0.205	0.355	0.475	0.45
AUC	0.698	0.294	0.623	0.571	0.595
P_{R100}	0.585	0.214	0.357	0.477	0.456
	Nordland Dataset				
VPR Technique	NetVLAD	HOG	CALC	AMOSNet	HybridNet
K	1	1	13	1	1
t_e (sec)	0.77	0.0043	0.351	0.36	0.36
t_m (sec)	0.155	1.62	0.563	0.536	0.613
t_{VPR} (sec)	0.925	1.624	0.914	**0.896**	0.973
Accuracy	**0.412**	0.023	0.143	0.162	0.186
AUC	0.733	0.036	0.39	0.132	0.214
P_{R100}	0.42	0.04	0.143	0.163	0.187

to achieve the same or better levels of performance as other CNN-based VPR techniques such as NetVLAD, AMOSNet or HybridNet.

This experiment concludes that, on Campus Loop dataset, CALC with a sequence length ($\mathbf{K}$) of 16 images can achieve better and faster place matching performance within the computational budget represented by the t_{VPR} of the single-frame-based implementation of NetVLAD. For this reason, we propose that the sequence-based implementation of CALC (with a sequence length of $\mathbf{K} = 16$ images) is selected as an alternative to the single-based implementation of NetVLAD ($\mathbf{K} = 1$) on this dataset, as presented in our experiment.

VI. CONCLUSION

To bridge the gap of lack of a systematic study on sequence-based filtering for visual route-based navigation, this paper has conducted an in-depth investigation on the benefits and trade-offs of sequence-based filtering on top of single-frame-based VPR methods. This analysis is performed on 4 public sequential VPR datasets, that pose difficulties in place matching (appearance changes, viewpoint variations etc), using a variety of widely used performance metrics, such as Performance-per-Compute-Unit (PCU). Sequential filtering is introduced into a number of contemporary single-frame-based VPR methods in order to present the findings. The results show the effects of various sequence lengths on performance boost and suitable combinations of different VPR techniques and sequence lengths are determined, taking into consideration the computational effects of sequential-filtering, for the best place matching performance in different scenarios.

This work uses a simple matching schema to highlight the benefits of using multiple images for VPR. A natural extension of this work is comparing different matching schema. While we demonstrated that VPR accuracy generally benefits from using a sequence of images to find a place, sequence matching has some more strict requirements than single-matching approaches. The most relevant requirement is in regards to the velocity of the traverses. If the velocity of the reference sequence if too different from that of the query, the matching might fail [1]. Thus, the analysis proposed in this paper could be extended to more complex sequence-based matching techniques to understand whether the trade-off between the sequence length, VPR performance and computational cost are affected by the matching method.

REFERENCES

[1] M. J. Milford and G. F. Wyeth, "SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 1643–1649.

[2] M. Milford, "Vision-based place recognition: How low can you go?" *Int. J. Robot. Res.*, vol. 32, no. 7, pp. 766–789, Jun. 2013.

[3] M. Chancán and M. Milford, "DeepSeqSLAM: A trainable CNN+RNN for joint global description and sequence-based place recognition," 2020, *arXiv:2011.08518*.

[4] O. Vysotska and C. Stachniss, "Lazy data association for image sequences matching under substantial appearance changes," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 213–220, Jan. 2016.

[5] M.-A. Tomia, M. Zaffar, M. J. Milford, K. D. McDonald-Maier, and S. Ehsan, "ConvSequential-SLAM: A sequence-based, training-less visual place recognition technique for changing environments," *IEEE Access*, vol. 9, pp. 118673–118683, 2021.

[6] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[7] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Computer Vision—ECCV 2006*, vol. 110. Berlin, Germany: Springer, Jul. 2006, pp. 404–417.

[8] A. C. Murillo, J. J. Guerrero, and C. Sagues, "SURF features for efficient robot localization with omnidirectional images," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3901–3907.

[9] E. Stumm, C. Mei, and S. Lacroix, "Probabilistic place recognition with covisibility maps," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 4158–4163.

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

[11] J. Košecká, F. Li, and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," *Robot. Auton. Syst.*, vol. 52, no. 1, pp. 27–38, 2005.

[12] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *Int. J. Robot. Res.*, vol. 21, no. 8, pp. 735–758, 2002.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.

[14] H. Jegou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[15] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.*, vol. 155, pp. 23–36, Oct. 2006.

[16] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[17] 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," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2136–2174, Jul. 2021.

[18] N. Sünderhauf and P. Protzel, "BRIEF-Gist—Closing the loop by simple means," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 1234–1241.

[19] W. T. Freeman and M. Roth, "Orientation histograms for hand gesture recognition," in *Proc. Int. Workshop Autom. Face Gesture Recognit.*, vol. 12, 1995, pp. 296–301.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.

[21] A. C. Murillo and J. Kosecka, "Experiments in place recognition using gist panoramas," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Sep. 2009, pp. 2196–2203.

[22] G. Singh and J. Kosecka, "Visual loop closing using gist descriptors in Manhattan world," in *Proc. ICRA Omnidirectional Vis. Workshop*, 2010, pp. 4042–4047.

[23] C. Mcmanus, B. Upcroft, and P. Newmann, "Scene signatures: Localised and point-less features for localisation," in *Proc. Robot., Sci. Syst. X*, Jul. 2014, pp. 1–9.

[24] 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 Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1835–1842, Apr. 2020.

[25] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3223–3230.

[26] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of ConvNet features for place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 4297–4304.

[27] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5297–5307.

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

[29] E. Pepperell, P. I. Corke, and M. J. Milford, "All-environment visual place recognition with SMART," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2014, pp. 1612–1618.

[30] M. Yang, J. Mao, X. He, L. Zhang, and X. Hu, "A sequence-based visual place recognition method for aerial mobile robots," *J. Phys., Conf. Ser.*, vol. 1654, no. 1, Oct. 2020, Art. no. 012080.

[31] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1–7.

[32] S. Garg and M. Milford, "Fast, compact and highly scalable visual place recognition through sequence-based matching of overloaded representations," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 3341–3348.

[33] E. Johns and G.-Z. Yang, "Feature co-occurrence maps: Appearance-based localisation throughout the day," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 3212–3218.

[34] P. Hansen and B. Browning, "Visual place recognition using HMM sequence matching," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 4549–4555.

[35] F. Lu, B. Chen, X.-D. Zhou, and D. Song, "STA-VPR: Spatio-temporal alignment for visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 4297–4304, Jul. 2021.

[36] L. G. Camara, C. Gäbert, and L. Přeučil, "Highly robust visual place recognition through spatial matching of CNN features," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 3748–3755.

[37] L. G. Camara and L. Přeučil, "Visual place recognition by spatial matching of high-level CNN features," *Robot. Auto. Syst.*, vol. 133, Nov. 2020, Art. no. 103625.

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

[39] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 404–417.

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Feb. 2004.

[41] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *Int. J. Robot. Res.*, vol. 30, no. 9, pp. 1100–1123, Jun. 2011.

[42] W. Maddern, M. Milford, and G. Wyeth, "CAT-SLAM: Probabilistic localisation and mapping using a continuous appearance-based trajectory," *Int. J. Robot. Res.*, vol. 31, no. 4, pp. 429–451, Apr. 2012.

[43] M. Agrawal, K. Konolige, and M. R. Blas, "CenSurE: Center surround extremas for realtime feature detection and matching," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2008, pp. 102–115.

[44] K. Konolige and M. Agrawal, "FrameSLAM: From bundle adjustment to real-time visual mapping," *IEEE Trans. Robot.*, vol. 24, no. 5, pp. 1066–1077, Oct. 2008.

[45] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Incremental vision-based topological SLAM," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2008, pp. 1031–1036.

[46] M. Waheed, M. Milford, K. McDonald-Maier, and S. Ehsan, "Improving visual place recognition performance by maximising complementarity," *IEEE Robot. Autom. Lett.*, vol. 6, no. 3, pp. 5976–5983, Jul. 2021.

[47] M. Waheed, M. Milford, K. McDonald-Maier, and S. Ehsan, "SwitchHit: A probabilistic, complementarity-based switching system for improved visual place recognition in changing environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022.

[48] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.

[49] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.
[50] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," 2014, *arXiv:1411.1509*.
[51] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[53] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from ConvNet for visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 9–16.
[54] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A holistic visual place recognition approach using lightweight CNNs for significant viewpoint and appearance changes," *IEEE Trans. Robot.*, vol. 36, no. 2, pp. 561–569, Apr. 2020.
[55] B. Ferrarini, M. J. Milford, K. D. McDonald-Maier, and S. Ehsan, "Binary neural networks for memory-efficient and effective visual place recognition in changing environments," *IEEE Trans. Robot.*, early access, Mar. 2, 2022, doi: 10.1109/TRO.2022.3148908.
[56] B. Ferrarini, M. Milford, K. D. McDonald-Maier, and S. Ehsan, "Highly-efficient binary neural networks for visual place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022.
[57] B. Arcanjo, B. Ferrarini, M. Milford, K. D. McDonald-Maier, and S. Ehsan, "An efficient and scalable collection of fly-inspired voting units for visual place recognition in changing environments," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 2527–2534, Apr. 2022.
[58] 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?" 2019, *arXiv:1904.07967*.
[59] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier, "Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions," 2019, *arXiv:1903.09107*.
[60] S. Skrede. (2013). *Nordland Dataset*. [Online]. Available: https://nrkbeta.no/2013/01/15/nordlandsbanen-minute-by-minute-season-by-season/

MIHNEA-ALEXANDRU TOMIȚĂ received the B.Sc. degree in computer science from the University of Essex, Colchester, U.K., in 2019, where he is currently pursuing the Ph.D. degree. He is also a part of the National Centre for Nuclear Robotics (NCNR), University of Essex.

His current research interests include computer vision, sequence-based filtering, deep learning, and SLAM.

MUBARIZ ZAFFAR received the B.E. degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2016, and the M.Sc. degree in computer science and electronic engineering from the University of Essex, U.K., in 2020. He is currently pursuing the Ph.D. degree with the Cognitive Robotics Department, TU Delft, where he is a part of the Intelligent Vehicles Group.

His research interests include computer vision and deep learning for autonomous robotics, visual place recognition and robot navigation, SLAM, and embedded systems.

Mr. Zaffar was a recipient of the South-Asian Helix Innovation Award, the DICE Foundation Innovation Award, the IET Present-Around-The-World Regional Awards, and the NUST High Achiever's Award.

BRUNO FERRARINI (Student Member, IEEE) received the M.Sc. degree in electronic engineering and the M.Sc. degree in ICT engineering from the University of Parma, Italy, in 2002 and 2006, respectively, and the M.Sc. degree (by Dissertation) in computer and electronic systems from the University of Essex, Colchester, U.K., in 2016, where he is currently pursuing the Ph.D. degree in computer science. His current research interests include machine learning, local image feature extraction, visual place recognition, and binary neural networks.

MICHAEL J. MILFORD (Senior Member, IEEE) received the Bachelor of Mechanical and Space Engineering degree and the Ph.D. degree in electrical engineering from The University of Queensland (QUT), Brisbane, QLD, Australia.

He is currently an Associate Professor and a Australian Research Council Future Fellow with QUT, and a Chief Investigator of the Australian Centre of Excellence for Robotic Vision. He was a Research Fellow on the Thinking Systems Project with the Queensland Brain Institute, until 2010, when he became a Lecturer with QUT. He conducts interdisciplinary research into navigation across the fields of robotics, neuroscience, and computer vision.

Dr. Milford was a recipient of an Inaugural Australian Research Council Discovery Early Career Research Award, in 2012, and became a Microsoft Research Faculty Fellow, in 2013.

KLAUS D. MCDONALD-MAIER (Senior Member, IEEE) received the Dipl.-Ing. degree in electrical engineering from the University of Ulm, Ulm, Germany, the M.S. degree in electrical engineering from the Ecole Suprieure de Chimie Physique lectronique de Lyon, Villeurbanne, France, in 1995, and the Ph.D. degree in computer science from Friedrich Schiller University, Jena, Germany, in 1999.

He was a Systems Architect on reusable micro-controller cores and modules with the Infineon Technologies AG.s Cores and Modules Division, Munich, Germany, and a Lecturer in electronics engineering with the University of Kent, Canterbury, U.K. In 2005, he joined the University of Essex, Colchester, U.K., where he is currently a Professor with the School of Computer Science and Electronic Engineering. His current research interests include embedded systems and system-on-a-chip design, security, development support and technology, parallel and energy efficient architectures, and the application of soft computing and image processing techniques for real-world problems.

Dr. McDonald-Maier is a member of the Verband der Elektrotechnik Elektronik Informationstechnik and the British Computer Society, and a fellow of the Institution of Engineering and Technology.

SHOAIB EHSAN (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology Taxila, Taxila, Pakistan, in 2003, and the Ph.D. degree in computing and electronic systems with a specialization in computer vision from the University of Essex, Colchester, U.K., in 2012.

He has extensive industrial and academic experience in the areas of embedded systems, embedded software design, computer vision, and image processing. His current research interests include intrusion detection for embedded systems, local feature detection and description techniques, image feature matching, and performance analysis of vision systems.

Dr. Ehsan was a recipient of the University of Essex Post Graduate Research Scholarship and the Overseas Research Student Scholarship. He is a winner of the prestigious Sullivan Doctoral Thesis Prize by the British Machine Vision Association.