Robust Visual Localization Across Seasons

Tayyab Naseer, Wolfram Burgard, and Cyrill Stachniss

***Abstract*—Localization is an integral part of reliable robot navigation and long-term autonomy requires robustness against perceptional changes in the environment during localization. In the context of vision-based localization, such changes can be caused by illumination variations, occlusion, structural development, different weather conditions and seasons. In this paper, we present a novel approach for localizing a robot over longer periods of time using only monocular image data. We propose a novel data association approach for matching streams of incoming images to an image sequence stored in a database. Our method exploits network flows to leverage sequential information to improve the localization performance and to maintain several possible trajectories hypotheses in parallel. To compare images, we consider a semi-dense image description based on HOG features as well as global descriptors from Deep Convolutional Neural Networks trained on ImageNet for robust localization. We perform extensive evaluations on a variety of datasets and show that our approach outperforms existing state-of-the-art approaches.**

I. Introduction

Monocular camera-based visual localization plays a vital role for navigation of autonomous vehicles. Robustly localizing a robot over longer periods of time in an environment that undergoes drastic perceptual changes due to changes in weather conditions, time of the day, or seasons is still a challenging problem. Various novel methods for robust place recognition have been proposed in the past including FAB-MAP2 [11], SeqSLAM [35], SP-ACP [39], FrameSLAM [1], and place-dependent feature learning [32]. Some of these approaches show impressive robustness to various changing conditions occurring due to different illumination or varying weather conditions. Most approaches aim to develop image descriptions that are repeatable over longer periods of time and enable long term visual localization.

This paper is an extension of our previous work [36] and addresses the problem of visual localization under large perceptual changes using image sequences collected along routes. Related methods have been developed over time that exploit sequential nature of the recorded reference image sequences such as SeqSLAM [35], SP-ACP [39], and RTPL [3]. These methods achieve robust localization under changing perceptual conditions due to day-night scenarios, different illumination and seasons. Often, trajectory-based approaches substantially reduce the false positive matches as compared to single image-based localization methods.

The key contribution of this work is a novel approach to visual place recognition under a large variety of perceptual changes. Our method successfully matches two image sequences *without* using any pose priors, for example, from GPS or robot odometry to normalize the robot speed. We do not perform explicit condition-based feature training for learning image descriptions. Our approach does not require any image pre-processing for exact image alignments. It can handle variable frame rates of the camera and different vehicle speeds in the database and localization routes. Our method also handles intra-trajectory loop closures and visits to new locations during the localization phase. It method achieves global localization without any pose initialization. We present a single framework which achieves all these objectives simultaneously. Fig. 1 shows successful matches of the same place across seasons and across day and night obtained with our approach.

Fig. 1: This figure shows the successful matches using our approach. On the left, we see the live image captured from the robot and on the right is the retrieved image from the database. It shows that our approach is robust to challenging day and night scenarios, foliage color changes and occlusions due to different seasons.

In this paper, we propose an approach to robust localization that does not rely upon either hand-crafted point features or complex global image descriptors plus additional constraints. Our method builds upon the semi-dense image description with HOG [13] descriptors as used by Naseer *et al.* [36]. Although this image representation is somewhat viewpoint dependent due to the tessellated gradient representation, it

Tayyab Naseer and Wolfram Burgard are with University of Freiburg, Germany. email:{naseer, burgard}@cs.uni-freiburg.de. Cyrill Stachniss is with University of Bonn, Germany. email:{cyrill.stachniss@igg.uni-bonn.de}

Fig. 2: Feature-based matching of the same place across seasons. In this case, SURF keypoints and their descriptions change drastically over seasons and lead to false correspondences.

provides better image matching performance than the point features. Recently, deep convolutional neural networks (DCNNs) have shown to outperform traditional feature-based approaches for various image recognition and classification tasks. Along with the hand-crafted features like SIFT [29] and HOG, we demonstrate that although the features from these complex neural networks provide more discriminative image matchings, it is still insufficient to achieve robust localization based only on the global image descriptor matchings. We build a data association graph that relates images from sequences retrieved in different conditions. The special node-transition connectivity enables us to compute multiple route hypotheses, deal with occlusions over short periods of time, and handle deviations from the previously taken route. We solve the visual place recognition problem by computing network flows in the association graph and generate multiple vehicle route hypotheses. By exploiting the specific structure of our graph, we solve this problem efficiently. We show that leveraging the sequential information using network flow provides high gain in the localization performance even when integrated with more robust image descriptions from DCNNs. Our extensive experimental evaluation suggests, that our method enables the vehicle to robustly localize across large perceptual changes in the environment based purely on the vision data and does not require any prior knowledge.

II. Related Work

Vision-based topometric localization has been studied extensively both in computer vision and robotics, see Garcia-Fidalgo and Ortiz [19] for a survey. Researchers have proposed robust feature-based approaches over the years for robot localization in similar environment appearances [11, 14, 1, 6, 18]. Biber *et al.* [7] deal with changing indoor environments by sampling laser maps at multiple timescales. Each sample of the map at a particular timescale is maintained and updated using the sensor data of the robot. This allows them to model spatio-temporal variations in the map. Kranjik *et al.* [26] use frequency spectra to model the dynamics of the object occurrences. The authors formulate it as a path planning algorithm and can perform efficient search for object localization using dynamic maps. Dymczyk *et al.* [17] propose an approach to summarize and update maps for long-term navigation by keeping a minimal number of landmarks in the memory for localization. Stachniss *et al.* [48], in contrast, aim at modeling different instances of typical states of the world using a clustering approach.

Although various approaches for large scale vision-based localization in dynamic indoor environments and similar perceptual conditions have shown promising results, localization under extreme variations in outdoor scenarios is still a hard and an unsolved problem. It has been recognized as a major obstacle for persistent autonomous navigation and has been addressed by different researchers [21, 11]. Many of the visual place recognition approaches rely on image matching by using features such as SURF [4] and SIFT [29]. Such feature-based algorithms work reliably for matching images that undergo rotation and scale variations but are susceptible to extreme perceptual changes caused by illumination changes, different weather conditions and seasons. Illumination changes as encountered during the period of a day degrades the performance of feature-based localization. Paton *et al.* propose an approach for learning color constant representation of RGB images and cope with illumination changes during a day [40]. It enables a robot to navigate autonomously through out the day. Researchers have also shown promising results for all-day localization by converting the images to illumination invariant color space [30, 31, 42].

Valgren *et al.* evaluate SIFT and SURF features in combination with geometric keypoint constraints for across-seasons image matching [51]. For our datasets, we found that both features do not match robustly, see Fig. 2 for an example. As a result, methods such as FAB-MAP2 [11] that require a reliable matching of such features tend to perform poorly. Kranjik *et al.* learn a sequence of comparisons for BRIEF [8] descriptors with evolutionary algorithms to localize a robot across seasons [25]. McManus *et al.* learns stable regions in images over days and months by leveraging big training data [32]. The approach learns distinctive visual elements in images and produces region detectors, which can be robustly associated across different illumination and weather conditions. In contrast to this, Carlevaris-Bianco *et al.* learn dynamics of feature point descriptions over different lighting conditions [9]. They generate training data by tracking feature points in time-lapse videos and use this training data to project feature descriptors to a lower dimensional space, which provides better discriminative power between descriptors under challenging lighting conditions.

Recently, featureless sequence-based SLAM (SeqSLAM) has shown a great improvement over feature-based global image localization [34, 35]. This approach achieves promising results for localization across day and night but does not explicitly address the problem of non-linear sequential matching and handling unseen places while revisiting the previously taken route. SeqSLAM can be seen as a continuous version of dynamic time warping (DTW) proposed by Sakoe and Chiba [45]. Our proposed method is in spirit similar to DTW with greater flexibility in non-linear transitions, non-overlapping trajectories and multiple loop closures.

Furthermore, there has been extensive research in the area of spatio-temporal video alignment methods [15, 16, 43]. Such methods aim at jointly optimizing the spatial image alignment and temporal frame correspondence of video sequences. Most

of these methods rely on point feature-based approaches to calculate relative homography between images. Unfortunately, these keypoints often change under large environmental changes (compare Fig. 2 for SIFT features). Diego et al. [15] leveraged GPS information and model the sequence alignment as a MAP inference problem. Our method does not assume any pose priors from other sensors. Ranganathan [41] proposes an algorithm (PLISS) to partition the image sequences for place categorization by integrating visual cues. PLISS shows impressive results in indoor place categorization where rooms have different appearance types. For an outdoor urban environment, where places have generally similar appearance types, the discriminating ability of PLISS to partition the image sequence has not been shown yet. Liu and Siegwart [28] propose an online inference method for topological place categorization using Dirichlet process mixture model and a light weight image descriptor. Although the authors concentrate on indoor scenes, such statistical models can also be investigated for place categorization in semi-structured outdoor scenarios. These methods mainly target indoor scene recognition or categorization and their performance has not been quantified for place recognition under adverse environmental conditions. Thus, it is unclear if they can be applied to the problem investigated in this work.

Neubert *et al.* propose to combine an appearance change prediction with a vocabulary-based method to predict how the visual word in the current scene would appear under changes [39]. For learning the appearance changes across seasons, an accurate image alignment is assumed. Johns *et al.* learn discriminative statistics on the co-occurrence of features over different conditions [24]. Their approach combines stable and discriminative features into one compact model from the data captured during different times of the day. In our experiments, however, we were unable to obtain such stable and discriminative features under the strong seasonal changes that we experienced. Instead of explicitly addressing the visual place recognition with extreme perceptual differences, Winston *et al.* associate different appearances, which they call as experiences, to the same place [10]. They localize in previously learned experiences and associate a new experience in case of a localization failure. At least during the setup phase, this requires some degree of place knowledge to assign a new experience to an existing place.

All of the aforementioned approaches use either handcrafted features or raw image intensities for image matching. Recently the feature representations from large Deep Convolutional Neural Networks (DCNNs) have shown to outperform the existing features for image classification and recognition tasks [27]. These networks are trained over millions of images for image classification and object detection tasks. A recent work by Sünderhauf *et al.* investigates the performance of DCNNs for the application of place recognition [49]. It evaluates the feature descriptions from DCNNs on various datasets to determine the impact of each layer of the place recognition performance. In another approach, the author proposed to combine region proposals and DCNNs to make the image matching more robust to viewpoint differences and appearance changes [50]. Neubert *et al.* proposed an approach to segment

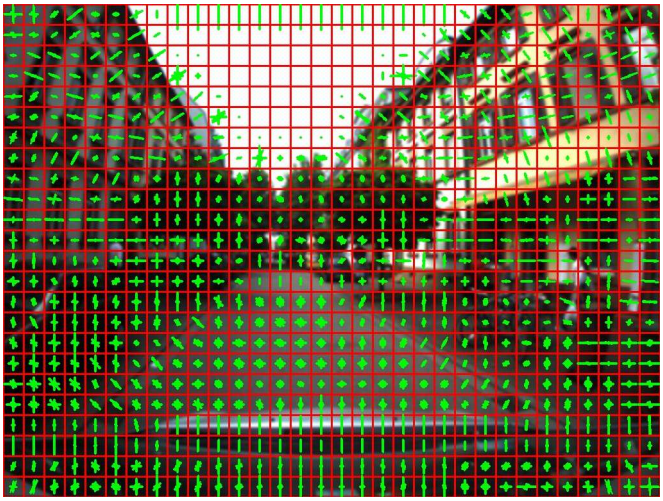

Fig. 3: We compute the HOG descriptor on the dense grid of 32×32 pixels over the entire image. The global description is the concatenation of all the cell descriptions which elevates the problem of unstable keypoint representation.

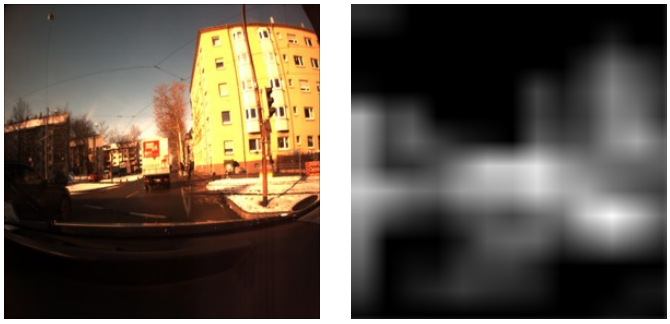

Fig. 4: This figure shows an image from the Freiburg dataset fed to the Alexnet and the corresponding feature map for one of the filters for conv3 layer. It shows that the feature map focuses on the structural part rather than non-salient parts like sky of the image.

patches based on multiscale superpixel grids instead of the bounding boxes to achieve better performance than the method of [50]. In this paper, we demonstrate that image matching from even such complex networks is inaccurate and leveraging non-linear sequential information can substantially boost the performance.

Related to our method, Vysotska *et al.* [55] build up a data association graph exploiting GPS information to simplify the process of finding a sequence of matching images in an offline fashion. For online localization, a search technique similar to A^* but with a non-admissible heuristic can be used to efficiently find the current best matching sequence [53, 52]. This approach can operate online but will commit in one hypotheses and will not return multiple parallel ones as the network flow approach presented in this paper. A further extension of these works, exploits hashing techniques to realize efficient relocalization after the vehicle has left the previously mapped area [54]. Robust data associations using our approach and robot odometry can be integrated into a visual SLAM framework to produce consistent trajectories across seasons [37]. Our approach uses network flows to con-

sider the sequential nature of the data in the individual routes. In other fields, network flows have been successfully used to address data association problems when tracking multiple people [56, 5]. Our proposed approach does not assume pre-processed images and does not require any sort of environment or condition specific feature learning. Naseer *et al.* [38] use a Markov localization framework with HOG descriptor-based image matching to show that sequential filtering can be performed online. This approach, however, might suffer from highly ambiguous false positives as it updates the state over the whole database instead of the constrained graph-connectivity as in our network flow approach.

III. Visual Route Matching Across Seasons

We define the set $\mathcal{D} = (d_1, \ldots, d_D)$ as the temporally ordered set of images that constitutes the visual map of places (the database) and $D = |\mathcal{D}|$. The set $\mathcal{Q} = (q_1, \ldots, q_Q)$ with $Q = |\mathcal{Q}|$ refers to the query sequence that was recorded in a different season or after a substantial scene change.

A. Matching Images

Vision-based localization for long-term navigation of autonomous robots is an important and challenging problem. The perceptional changes in the environment can be caused by different times of the day, structural changes, different weather conditions and seasons. View-point and scale variance in most real-world applications while revisiting the same place makes the problem harder. Traditional approaches extract keypoints on the images and compute a hand-crafted feature descriptor for that keypoint [33]. This is a sparse representation of the image contents and highly depends on the keypoints repeatability. The keypoints do not remain stable over large time lags and the description of the keypoints changes dramatically which leads to false keypoint correspondences as show in Fig. 2. We present a novel semi-dense image description based on HOG descriptors and also discuss the advantages of recently introduced global features from Deep Convolutional Neural Networks (DCNNs).

HOG: Instead of a sparse keypoint-based image description, we propose a semi-dense global image description. To achieve this we compute HOG descriptors on a dense grid of 32×32 pixels over the whole image of size 1024×768 pixels. This dense representation allows viewpoint variance over an image patch of 32×32 pixels as HOG describes a region by accumulating the gradient regardless to its location. The overall image descriptor $\mathbf{h}$ is a vector composed of the concatenation of all the histograms of gradients computed on all the cells as shown in Fig. 3.

DCNNs: Recently, deep networks have gained great importance in the computer vision community for various object detection, image recognition and classification tasks. DCNNs learn neuron weights over all stages of the network when trained on a large number of labeled images. They consist of convolutional layers in early stages while later layers provide generic representation of the images for classification tasks. In our application, the task of the vehicle is to recognize

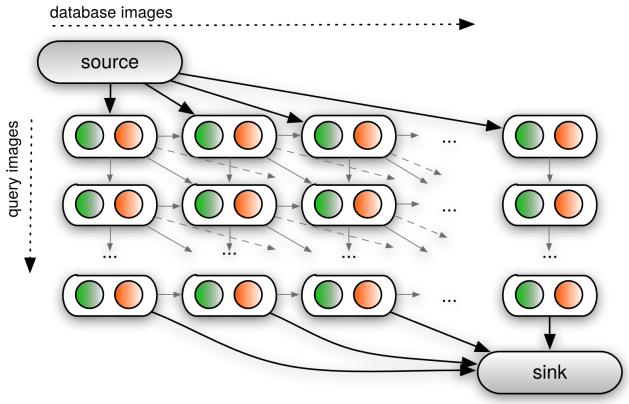

Fig. 5: Illustration of the data association graph.

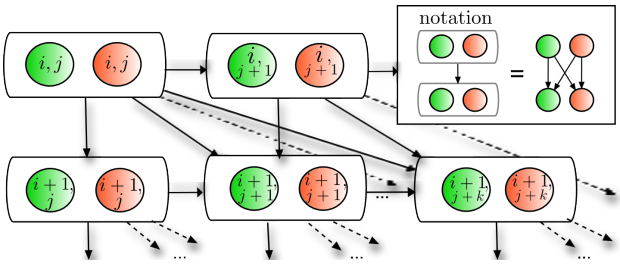

Fig. 6: Illustration of the connections $\mathcal{E}^a$ and $\mathcal{E}^b$ between matching (green) and hidden (red) nodes. The green rounded rectangles are only used for illustration: an edge between two rounded rectangles means that all nodes contained in the first rectangle are connected via a directed edge to all nodes in the second rectangle.

the similarity between the places so we extract feature representation from early stages of the network. For this, we use the Alexnet [27] architecture pre-trained on the ImageNet database [44] inside the Caffe framework [23]. The conv3 layer of the Alexnet architecture has been shown to be robust against perceptual changes [49], which is why we also use it for our application of place recognition over longer periods of time. We crop images to the size of 256×256×3 before feeding them to the deep network and extract features from the 3^{rd} convolutional layer which has 384 filters with spatial dimension of 13×13, resulting in a global descriptor size of $64,896$. A feature visualization in such a setting is shown in Fig. 4. The global descriptors are normalized and matched according to Eq. (1). The idea is to show that, although the feature representation from the complex neural networks boosts the localization performance, image matching is not perfect and high gain is achieved by leveraging the sequential information.

We compute the similarity between image $q_i \in \mathcal{Q}$ and $d_j \in \mathcal{D}$ by the cosine distance of the two image descriptors, respectively $\mathbf{h}_{q_i}$ and $\mathbf{h}_{d_j}$:

$$c_{ij} = \frac{\mathbf{h}_{d_j} \cdot \mathbf{h}_{q_i}}{\|\mathbf{h}_{d_j}\| \|\mathbf{h}_{q_i}\|}, \tag{1}$$

where $c_{ij} \in [0, 1]$ and $c_{ij} = 1$ indicates a perfect match. The matching matrix $\mathbf{C}$ has a size of $Q \times D$ and consists of all c_{ij}, i.e., the cosine distances between all images of

$\mathcal{Q}$ and $\mathcal{D}$, computed according to Eq. (1). Normalization of these similarity scores provide more distinctive values [38]. We normalize these scores by the mean of scores over each column according to Eq. (2).

$$\hat{c}_{ij} \quad = \quad c_{ij} \left(\frac{1}{Q} \sum_{i=1,\dots,Q} c_{ij} \right)^{-1} \tag{2}$$

This normalization reduces the ambiguities arising from the confusing database images which could match against most of the query images.

In our approach, the network is a graph $\mathcal{G} = (\mathcal{X}, \mathcal{E})$, where $\mathcal{X}$ are the nodes and $\mathcal{E}$ the edges. We denote the quantity of flow generated by the source node as $F \in \mathbb{N}$.

B. Building the Data Association Flow Network

A standard approach to image-based localization returns for a query image in $\mathcal{Q}$ the best matching image in $\mathcal{D}$ according to the matching matrix $\mathbf{C}$. Due to the visual change across seasons, a best-match-strategy in $\mathbf{C}$ typically results in a poor localization performance. In this paper, we leverage that $\mathcal{Q}$ and $\mathcal{D}$ consist of *image sequences* that are recorded on a robot or vehicle. As a result, locations are visited progressively and images are not in random order. The matching patterns in the matching matrix $\mathbf{C}$ reflect the temporal information of both sequences. Our approach exploits the sequential nature of data but does not assume that every image in $\mathcal{Q}$ has a matching counterpart in $\mathcal{D}$. We consider sequences that can start and stop at any position in the query and database set. Both sets might be composed of images that have been recorded at different framerates or while traveling at different speeds.

In this paper, we propose to solve the visual route localization problem by building a flow network and computing its minimum cost flow. The minimum cost flow problem consists of determining the most cost-effective way for sending a fixed amount of flow through a network [2]. The flow network is a directed graph with at least one source node and one sink node. The source node is the one that produces flow and the sink node is the one that consumes flow. To each edge, we associate a cost w and a capacity r. A source node is connected by only outgoing edges, a sink node by only ingoing edges. The capacity defines the number of units that can flow over an edge. Our idea is to build a flow network to model the possible matches between $\mathcal{D}$ and $\mathcal{Q}$. A minimum cost flow algorithm finds a set of paths that connect the source to the sink minimizing the path cost while transporting the specified flow to the sink. Those paths represent multiple hypotheses about the correct image matching and the estimation of the vehicle route. In order to match complex temporal sequences that include loops, we introduce special nodes to allow solutions that contain partially matched sequences.

1) Nodes: The set $\mathcal{X}$ contains four types of nodes: the *source* x^s, the *sink* x^t, the *matching nodes* x_{ij}, and so-called *hidden nodes* $\breve{x}_{ij}$. The node x^s is the node that creates all the flow F and x^t is the only sink that consumes it. A node x_{ij} represents a match between the i-th image in $\mathcal{Q}$ and the j-th image $\mathcal{D}$, i.e., that both images are from the same location. There exists a hidden node $\breve{x}_{ij}$ for each matching node x_{ij}. The hidden nodes represent "non-matches" between images and such nodes allow for paths even though the image pairs cannot be matched. These nodes are traversed during short temporal occlusions or non-matching sequences that occur when the robot deviates from the original route.

2) Edges: The edges in $\mathcal{G}$ define the possible ways of traversing the graph from the source to the sink. Fig. 5 illustrates the connectivity of our graph. We define four types of edges in $\mathcal{E} = \{\mathcal{E}^s, \mathcal{E}^t, \mathcal{E}^a, \mathcal{E}^b\}$. The first set $\mathcal{E}^s$ connects the source to a matching node or to a hidden node:

$$\mathcal{E}^s \quad = \quad \{(x^s, x_{1j}), (x^s, \breve{x}_{1j})\}_{j=1,\dots,D} \tag{3}$$

The set $\mathcal{E}^s$ models that the first image of $\mathcal{Q}$ can be matched with any image in the $\mathcal{D}$ via the matching nodes or that no match is found via the hidden nodes. The second set of edges, $\mathcal{E}^t$, represents all the connections that go to the sink:

$$\mathcal{E}^t \quad = \quad \{(x_{Qj}, x^t), (\breve{x}_{Qj}, x^t)\}_{j=1,\dots,D} \tag{4}$$

The sink can be reached from any of the matching nodes x_{Qj} and from the corresponding hidden nodes $\breve{x}_{Qj}$ with $j = 1, \dots, D$. This models the matching or non-matching of the last query image.

The set $\mathcal{E}^a$ of edges establishes the connections between the matching nodes as well as between the hidden nodes. For clarity it has can be divided into forward edges $\mathcal{E}^{af}$ and horizontal edges $\mathcal{E}^{ah}$. The set $\mathcal{E}^{af}$ connects nodes from the current query image to the next and $\mathcal{E}^{ah}$ allows the vehicle to stop in the database for the same query image

$$\mathcal{E}^{af} \quad = \quad \{(x_{ij}, x_{(i+1)k}), (\breve{x}_{ij}, \breve{x}_{(i+1)k})\}_{\substack{i=1,\dots,Q \\ j=1,\dots,D \\ k=j,\dots,(j+K) \\ k \leq D}} \tag{5}$$

$$\mathcal{E}^{ah} \quad = \quad \{(x_{ij}, x_{i(j+1)}), (\breve{x}_{ij}, \breve{x}_{i(j+1)})\}_{\substack{i=1,\dots,Q \\ j=1,\dots,D-1}} \tag{6}$$

$$\mathcal{E}^{a} \quad = \quad \mathcal{E}^{af} \cup \mathcal{E}^{ah}, \tag{7}$$

where $k = j, \dots, (j + K)$. These edges allow for finding sequences of matching images or sequences of unmatched query images respectively. Finally, the last set $\mathcal{E}^b$ of edges connects hidden and matching nodes:

$$\mathcal{E}^b \quad = \quad \{(x_{ij}, \breve{x}_{(i+1)k}), (\breve{x}_{ij}, x_{(i+1)k})\}_{\substack{i=1,\dots,Q \\ j=1,\dots,D, \\ k=j,\dots,(j+K) \\ k \leq D}} \tag{8}$$

The edges in $\mathcal{E}^b$ are the ones that are traversed when the sequence is not continued with the children of a node. Edges in $\mathcal{E}^b$ are the ones that are traversed when a matching is found again so that the matching sequence can continue. See Fig. 6 for an illustration of the edges in $\mathcal{E}^a$ and $\mathcal{E}^b$. As a design decision, there are no edges connecting nodes back in time, mainly for constraining the search space. However, this is not a limiting factor: loops in the route can be found by solving the flow network when $F > 1$.

The value of K specifies the number of considered path hypotheses exiting from each node: the fan-out from a vertex defines which of the subsequent images can be concatenated to a path. Values for $K > 1$ allow for matching sequences

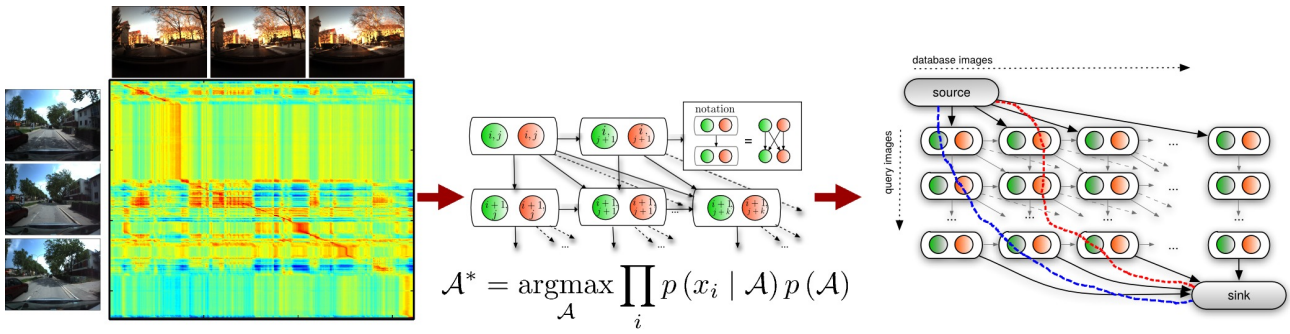

Fig. 7: This figure shows the pipeline of our approach. We extract robust image features from image sequences to compute a similarity matrix. We then build a data association graph over the matrix and leverage the sequential information using network flows to compute multiple path hypothesis for vehicle localization.

recorded at different vehicle speeds or in case of different camera framerates. An edge between nodes (i, j) and $(i+1, j)$ models a vehicle that does not move in $\mathcal{Q}$ and an edge between (i, j) and $(i, j+1)$ models that a vehicle does not move in $\mathcal{D}$. In our implementation, we use $K = 4$, which, according to our experience in our experiments, seems to be a sufficient value for typical city-like navigation scenarios. Edges connected to hidden states capture the fact that the corresponding images cannot be matched (due to strong changes, occlusions, etc.), but allow the path to continue through some hidden nodes. The hidden nodes can also be used to terminate a matching sequence without terminating the overall localization process. This is important to handle situations in which the vehicle temporarily deviates from the route taken during mapping. Thanks to this graph design, $\mathcal{G}$ is a directed acyclic graph (DAG).

3) Edge Costs and Capacity: The cost of an edge connected to a matching node x_{ij} is $w_{ij} = \frac{1}{\hat{c}_{ij}}$, where $\hat{c}_{ij}$ is computed according to Eq. (2). In the case in which the edge is connected to an hidden node, the weight is constant, $\breve{w} = W$. We determined W this parameter experimentally by using a precision-recall evaluation. In addition to that, we set the weight of the edges in $\mathcal{E}^t$ and $\mathcal{E}^{ah}$ to 0.

All edges that interconnect the hidden nodes have a capacity $r = F + 1$ so that they can be considered for usage for each unit of flow. All the other edges have a capacity of $r = 1$ so that they can be only used once. The path resulting from the minimum cost flow on $\mathcal{G}$ corresponds to the best data association between $\mathcal{D}$ and $\mathcal{Q}$.

C. Minimum Cost Flow For Vehicle Localization

In this section, we provide a probabilistic interpretation of our solution for solving this problem. Without loss of generality, we present a formulation for $F = 1$. We define the ordered set $\mathcal{A} = (x^s, x_{a_1}, \ldots, x_{a_A}, x^t)$ where x_{a_i} is a simplified notation indicating a vertex in $\mathcal{X}$. The sequence $\mathcal{A}$ is a route hypothesis, i.e., a sequence of matched images between seasons. It contains only vertices that can be connected with the edges presented in the previous section. Each sequence starts at the source and ends at the sink. For finding the best matching visual route, we find the optimal sequence $\mathcal{A}^*$ with a maximum a posteriori approach:

$$\begin{aligned} \mathcal{A}^* &= \underset{\mathcal{A}}{\operatorname{argmax}}\, p(\mathcal{A} \mid \mathcal{X}) \\ &= \underset{\mathcal{A}}{\operatorname{argmax}}\, p(\mathcal{X} \mid \mathcal{A})\, p(\mathcal{A}) \\ &= \underset{\mathcal{A}}{\operatorname{argmax}}\, p(\mathcal{A}) \prod_i p(x_i \mid \mathcal{A}) \end{aligned} \tag{9}$$

We consider all the x_i to be conditionally independent given $\mathcal{A}$ and define the prior $p(\mathcal{A})$ as

$$p(\mathcal{A}) \;=\; p_s\, p(x_{a_2} \mid x_{a_1}) \ldots p(x_{a_A} \mid x_{a_{A-1}}) p_t \tag{10}$$

where p_s and p_t are the priors associated to the source and sink. The term $p(x_{a_{i+1}} \mid x_{a_i})$ is proportional to $c_{a_{i+1}}$. We define the likelihood $p(x_i \mid \mathcal{A})$ of x_i being part of $\mathcal{A}$ as

$$p(x_i \mid \mathcal{A}) = \begin{cases} 1/Q & \text{if } x_i \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

To search for the best solution of Eq. (9), we use a minimum cost flow solver. An efficient implementation for minimum cost flow is the one of Goldberg and Kennedy [22], which has complexity $\mathcal{O}\left(|\mathcal{X}|^2 |\mathcal{E}| \log |\mathcal{X}|\right)$. In our context, this is expensive as typical problems consist of hundreds or thousands of images. Note that in the special case of $F = 1$, finding a minimal cost flow is equivalent to find the shortest path.

To solve this problem efficiently, we exploit the specific structure of the graph. Our graph $\mathcal{G}$ is a DAG with non-negative edges and each edge has either a capacity $r = 1$ or $r = F + 1$. This means that all paths through the matching nodes found by the minimum cost network flow consist in *different* paths. Given these restrictions, we formulate an equivalent solution with a substantially smaller computational complexity. Computing a shortest path in a DAG with a single source can be done by topological sorting in $\mathcal{O}(|\mathcal{X}| + |\mathcal{E}|)$, which is even more efficient than Dijkstra's or the Bellman-Ford algorithm. Note that, in our case, $|\mathcal{E}|$ depends linearly with respect to $|\mathcal{X}|$. For depleting all flow of the source node, we repeat this procedure F times. This leads to an overall complexity of $\mathcal{O}(F\,|\mathcal{X}|) = \mathcal{O}(F \cdot Q \cdot D)$.

Each execution of the solver leads to a loop-free path of the vehicle, as a consequence of our graph connectivity. The flow F controls the maximum number of vehicle path hypotheses

Fig. 8: The trajectories of all the sequences of *FAS* dataset are visualized in a satellite image. The vehicle route is shown in yellow. *Left:* The first sequence recorded in May 2012 with the trajectory length of 10 km. *Middle:* The trajectory traversed in December 2012. *Right:* The trajectory traversed in May 2015, which includes different routes because of road construction and traffic jams.

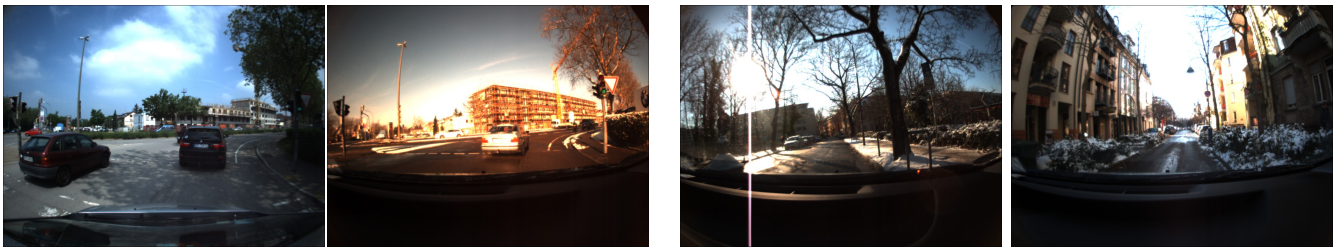

Fig. 9: The Freiburg dataset captures significant perceptual and structural changes over the span of 3 years. This figure shows one such scenario (left) where a new building was constructed within the period of five months between the visits.

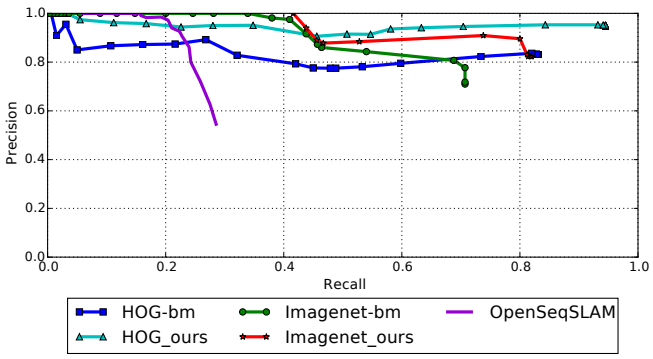

Fig. 10: Sequential information using our non-linear network flow approach boosts the localization performance outperforms the best match strategy for all the feature descriptors.

that are found in the graph. As there are at most F iterations, the system returns the F best paths. In this way, we are able to report sequences that include up to F traversals of the same loop. The parameter F is either set beforehand by limiting the search to F possible solutions or by repeating the search until the computed path is dominated by hidden nodes (non-matching events).

IV. Experimental Evaluation

We evaluated our approach on a variety of datasets which manifests the generalization of our approach over various perceptual conditions. We furthermore illustrate that it outperforms two state-of-the-art methods, namely FABMAP2 and SeqSLAM. For the evaluation, we recorded datasets by driving through a city with a camera-equipped car during different seasons including summer and winter. We collected image data while driving a distance of ~ 55 km in Freiburg city, Germany, overlayed on Google Maps Fig. 8. The Freiburg sequences contain between $5,392$ and $30,790$ images. No rectification, cropping, or other preprocessing has been applied to the images. In the following subsection, we discuss in detail the introduced dataset defined as *Freiburg Across Seasons (FAS)*. Furthermore, we evaluated our approach on several publicly available datasets such as the VPRiCE-dataset[1], the Nordland dataset[2], the NewCollege Dataset[47] and the Gardenspoint-Walking dataset [3]. The datasets include, viewpoint variances, illumination changes and extreme seasonal variations. The datasets also exhibit visits to new places, overlapping routes, different camera frame rates and typical real world driving maneuvers affecting the speed of the vehicle.

A. Dataset: Freiburg Across Seasons (FAS)

In this subsection, we discuss the introduced dataset which captures the longterm perceptual changes across a span of 3 years. We recorded the image sequences with a forward facing bumblebee stereo camera mounted on a car. During summer, the camera was mounted outside the car where as during winters the camera was inside the car as it can be observed in Fig. 1. The image sequences are recorded at relatively low frame rates of 1 Hz and 4 Hz. All the images have a resolution of 1024×768 (width$\times$height) and are JPEG compressed. We do not perform any preprocessing on

[1] https://goo.gl/R0QYU2
[2] https://nrkbeta.no/2013/01/15/
[3] https://goo.gl/tqmWyq

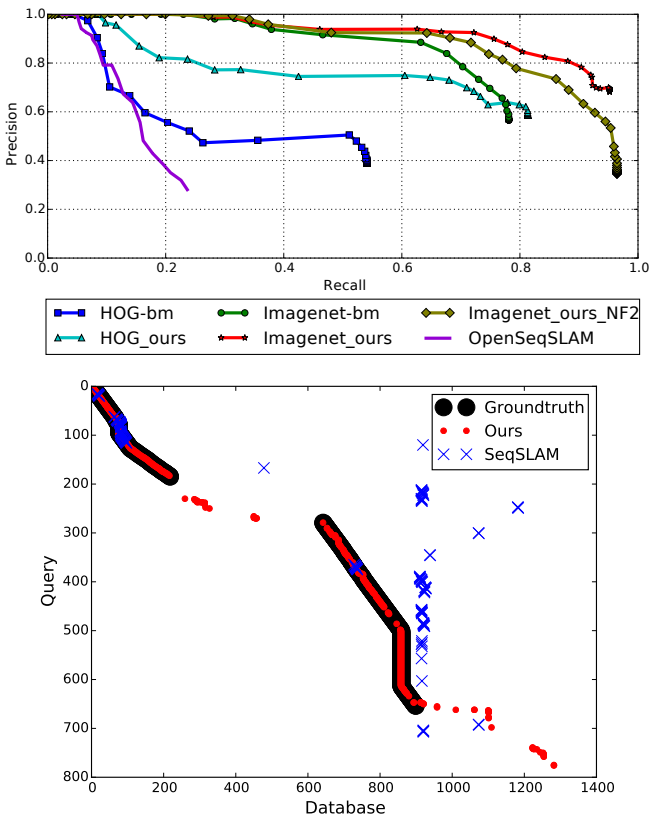

Fig. 11: *Top:* Our approach achieves higher average precision and outperforms best-match strategy and OpenSeqSLAM. The network flow solution with horizontal edges achieves better precision but suffers from a lower recall than the solution with two flows without horizontal edges (NF2). *Bottom:* Our approach successfully estimates route of the vehicle, even though it takes different routes in both the sequences with overlapping places. We take advantage of multiple flows in this example to estimate the complete route as single flow does not retrieve the whole path because of the constrained transition model.

the recorded images. We provide stereo images for the all the recorded sequences which can be further used to extract the depth information of the scene. All the images are geo-tagged with the corresponding GPS positions (which have been recorded with an inexpensive GPS sensor). We recorded the first sequence in May 2012[4] covering a distance of $10\,\mathrm{km}$. It contains 6915 images recorded at $1\,\mathrm{Hz}$. While driving through the city, we encountered all the natural driving maneuvers. The next sequence was recorded in Winter 2012[5] during the month of December to capture the large perceptual changes over these months. This sequence covers a distance of $50\,\mathrm{km}$. It contains $30,790$ images recorded at $4\,\mathrm{Hz}$. The routes were obstructed by road constructions, hence the exact overlap between the trajectories was not possible. Fig. 9 shows the snow covered roads, bare trees, strong sun glare and wet roads. Fig. 9 also shows an example of structural change that is captured during the recording of these sequences. All such perceptual variations are captured in our datasets making it valuable for the evaluation of vision-based methods for longterm localization. We recorded the third sequence in Summer 2015[6] during the month of May. The trajectory of the sequence is shown in Fig. 8 (*Right*) and it contains $5,392$ images. We recorded two sequences that captured summer season in May 2012 and May 2015, therefore we define them as *Localization-1* and *Localization-2* respectively. The sequence recorded in Winter 2012 is defined as the *Mapping* sequence. We provide ground truth for all the localization sequences with reference to the *Mapping* sequence. All the images have been hand-labeled after manual visual inspection for the ground truth matching. We rank the images based on their GPS positions and then manually select all the images from the mapping sequence which correspond to the same place as from the localization sequence. *Localization-2* has $4,477$ images that correspond to the same place as in the *Mapping* sequence. This sequence provides a denser ground truth than *Localization-1* which has $3,656$ images representing same places. In total, we provide ground truth matchings for $8,133$ images[7]. This makes it one of the largest datasets that capture large perceptual changes in urban driving scenarios over multiple years.

B. Variable Vehicle Speeds

The first experiment is designed to show the effectiveness of our approach under different vehicle speeds. It consists of 676 summer images and 322 winter images from the Freiburg dataset. We furthermore demonstrate that our approach generalizes to different speeds of the vehicle while recording both database and query sequences. The vehicle traveled the same route with no visits to new places and no intra-trajectory loops. Our approach matches the image sequences from both the runs at a higher precision and recall. It successfully estimates the route of the vehicle although it had different speeds while recording the datasets and stopped couple of times at traffic signals. In this experiment HOG features combined with the sequential information performs the best. The Alexnet-bm do not achieve higher recall than HOG features, which in turn leads to lower performance even when integrated with the sequential information as shown in Fig. 10. We provide an extensive quantitative evaluation, further addressing the performance of DCNN features from different architectures and handcrafted features in IV-K.

C. Visits to New Places

The second experiment is designed to show that our approach is also robust to new visits of the vehicle while recording the localization sequence i.e there does not exist a match for every query image in the database. In this case, our network flow approach uses the special hidden nodes to continue the sequence while traversing the new places. The dataset consists of a subset of the Freiburg dataset and comprises of 781 summer images and $1,328$ winter images. In this experiment we also highlight the effect of using multiple flows and horizontal edges in the data association graph. In this example, the path can be retrieved either by adding horizontal edges or by multiple flows. Fig. 11 shows that the solution

[4] goo.gl/1Jf3kI [5] goo.gl/AvZvjc

[6] goo.gl/Y2I6CI [7] goo.gl/PIZIvz, goo.gl/GDkmMq

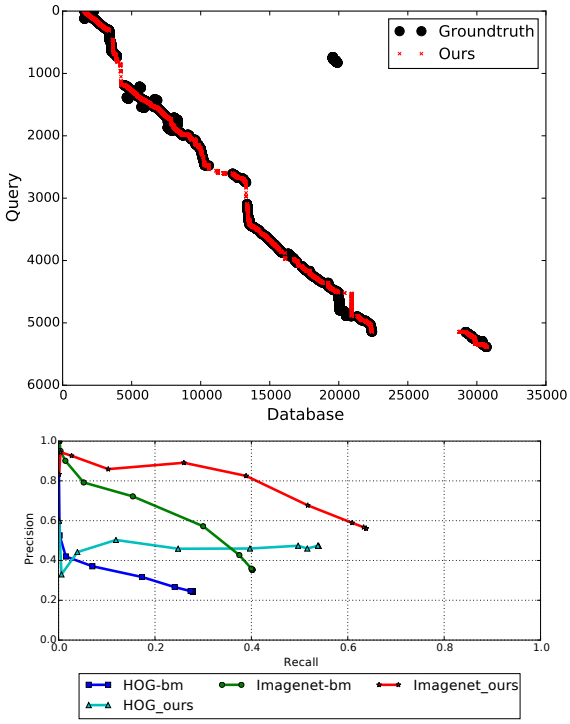

Fig. 12: *Top:* Our approach successfully estimates the path on the Freiburg dataset. It shows that our approach generalizes to matching large trajectories across seasons. The estimated path shows that our approach handles multiple stoppages both in query and the database, new traversals along with non-linear vehicle speeds. *Bottom:* The performance curve clearly shows that the non-linear sequential information adds significant gain over the best match using deep networks as well.

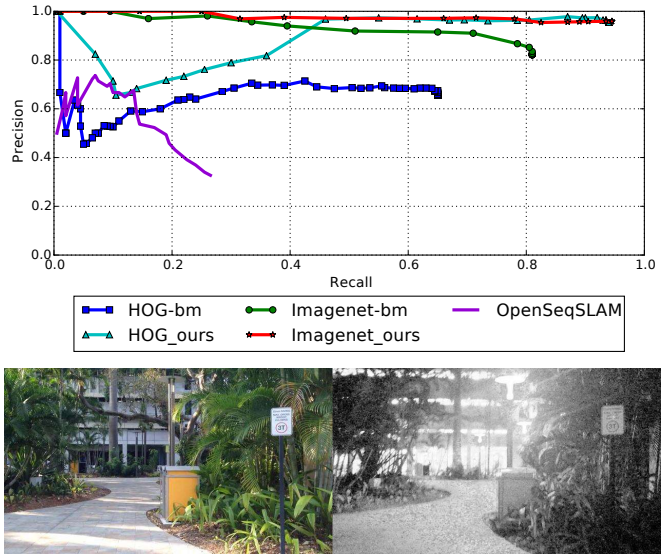

Fig. 13: *Top:* The precision recall curve exhibits the advantage of leveraging sequential information even with weaker hand-crafted descriptors like HOG. Although the best-match strategy using DCNN descriptors outperform HOG, with the sequential information both the descriptors achieve a precision and recall. *Bottom:* It shows pair of images successfully matched using our approach. It can be seen that our approach is robust to challenging conditions like noisy night images.

with multiple flows achieve higher recall but lower precision. It is because in the second flow the solution picks up more false positives which are avoided by introducing the horizontal edges. Horizontal edges in the graph enable to estimate the path of the vehicle in a single flow at the cost of missing some true matches. The horizontal edges can only help to reduce the number of flows in the case of a loop-less trajectory.

D. Scalability

The third experiment is designed to emphasize the scalability of our approach. In this experiment, we match the full trajectories from summer and winter season recorded in May 2015 and December 2012 respectively. We recorded the dataset while driving a 50 km long trajectory in Freiburg, Germany. It consists of $30,790$ winter images and 5,392 summer images. The results of our approach are shown in Fig. 12. For such large trajectories we achieve high gain in the localization performance by leveraging the sequential information using our network flow approach. The best-match strategy with the CNN features achieve 28% recall at 60% precision while with network flow we achieve 59% recall at 60% precision which is an improvement of 110%.

E. Generalization to Day-Night Scenarios

The next experiment is designed to show that we can also address changes that result from operating during day vs. night time and we used the GardenspointWalking dataset for this purpose. The dataset comprises of low resolutions and noisy images. It highlights the fact that our approach does not depend on high resolution images and results in robust localization not only across seasons but also across day and night. It also shows the effectiveness of the sequential gain along with the choice of hand-crafted features and the features from CNN. This sequence neither contains visits to new locations nor loops so all the images have one-to-one correspondence. We achieve 70% precision at 60% recall for HOG descriptors and 91% precision at 60% recall for CNN features as shown in Fig. 13. Interestingly, in this case most of the gain comes from adding the sequential constraints, where both HOG and CNN achieve 94% recall at 95% and 96% precision respectively. This experiment shows that in particular scenarios our network flow approach can even boost the performance of hand-crafted features by a great deal even if the descriptor-based best matches are not highly discriminative.

F. Non-Urban Environments

The next experiment is designed to exhibit the generalization of our approach to localize in non-structural environments under extreme seasonal variations. We evaluated our approach on the Nordland dataset to support the claim. This dataset contains sequence of images captured from a camera mounted on a train and covers a 728 km long trajectory of the mountainous areas through Norway. The images in top row of Fig. 14 shows the extreme seasonal changes across the two trajectories along the mountains. For descriptor-based best matches the CNN features outperform HOG features by 117% but this gain is reduced drastically when sequential

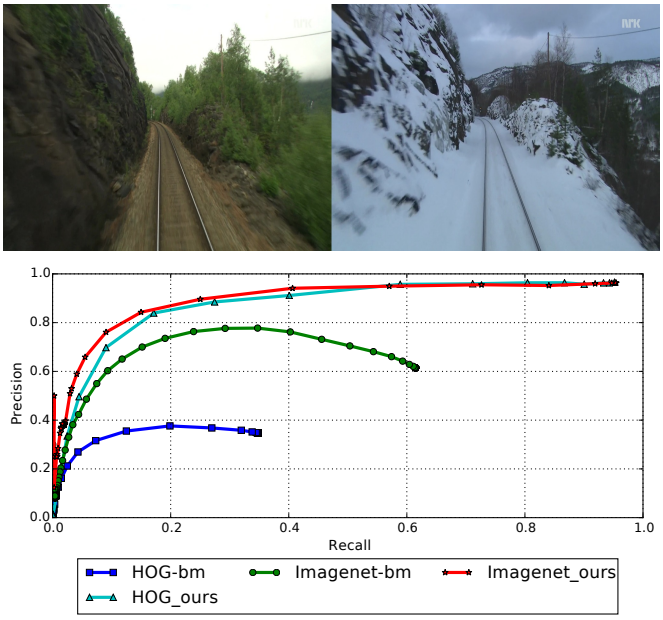

Fig. 14: *Top:* Our approach successfully matches this image pair from the Nordland dataset captured in two different seasons which demonstrates that the proposed approach generalizes to non-urban environments as well. *Bottom:* The Nordland dataset is a special case where the images are captured with a camera mounted on a train which follows exactly the same route every time. Although, the image description from DCNNs outperform the semi-dense HOG based matchings, the large performance boost comes from leveraging the sequential constraints. This dataset is a special case of a linear trajectory so even the HOG based image matchings when combined with the sequential information provide robust localization in this particular case.

information as added. Both the feature descriptions with sequential constraints achieve 95% recall at 96% precision in this case. The sequential gain is dramatically high because of the linear nature of the train trajectories which is a special case.

G. Heterogeneous Trajectories

The next experiment is designed to show the performance of our approach on a benchmark dataset to allow for better comparisons. Thus, we used the publicly available VPRiCE dataset using within the past years for benchmarking purposes. This dataset is challenging as it comprises of sequences with different season appearances, extreme view point variations (images captured from a bike and a bus of the same place), noisy low resolution images captured from a bike across day and night. It also includes images for bidirectional loop closure which have been ignored for precision recall calculation in our current setup as it is not the scope of our paper. The results on this challenging dataset are shown in Fig. 15. The performance of OpenSeqSLAM on this dataset is relatively better than the other datasets because half of the dataset consists of the images from Nordland dataset. As these images are pixel aligned without any viewpoint variance and linear trajectory, OpenSeqSLAM performs relatively better. Whereas, our approach retrieves matches from other sub-sequences as well which increases the recall and we achieve 74% recall with 81% precision and outperforms existing methods in this case which illustrates that our approach is robust to multiple perceptual changes and not tuned for a single condition change.

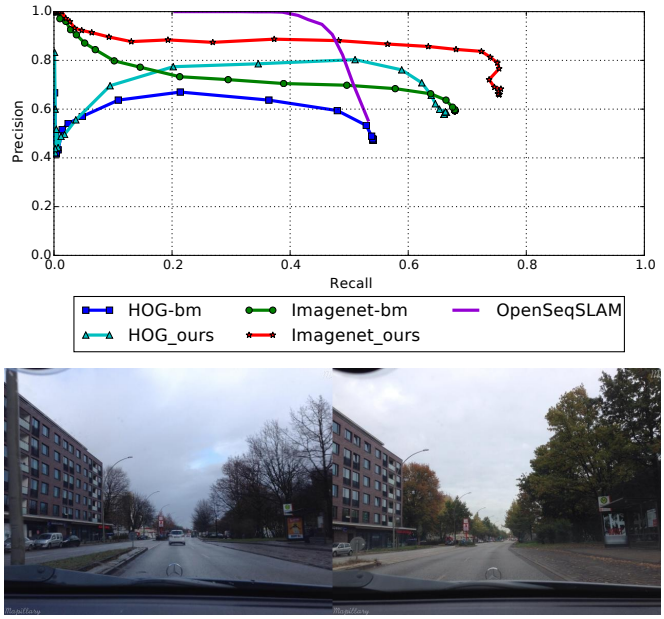

Fig. 15: *Top:* Leveraging the sequential information boosts the localization accuracy on the VPRiCE-dataset. Even the complex feature descriptions from deep nets do not provide exact image similarities and the temporal information helps in eliminating false positives which leads to better precision. *Bottom:* This pair of images show a successful match using our approach in one of the challenging scenarios where the two images are captured in different seasons and from different lanes of the road.

H. Intra-Trajectory Loops

This experiment is designed to illustrate the advantage of using multiple flows in our network flow approach. For this experiment we use the NewCollege [47] dataset. This dataset contains multiple loop closures in the same trajectory, which are not retrievable with a single flow. We show that using multiple flows increases the recall in this case as shown in Fig. 16. Using two and three flows the maximum recall is increased from 51% to 61% and 79% respectively.

I. Feature-based Localization

We also compare our method to FABMAP [12], a successful state-of-the-art approach for feature-based visual localization. For this experiment, we evaluated both OpenFABMAP2 and our approach on a sub-sequence of the Freiburg dataset. It consists of $1,213$ summer images and 596 images from the winter dataset. It has variable vehicle speeds and visits to new places. For the comparison to FABMAP2, we used the OpenFABMAP2 implementation by [20] with its default parameter settings. The original binary version of FABMAP2 [11] provided by the Oxford Mobile Robotics Group performs similar to OpenFABMAP2 . OpenFABMAP2

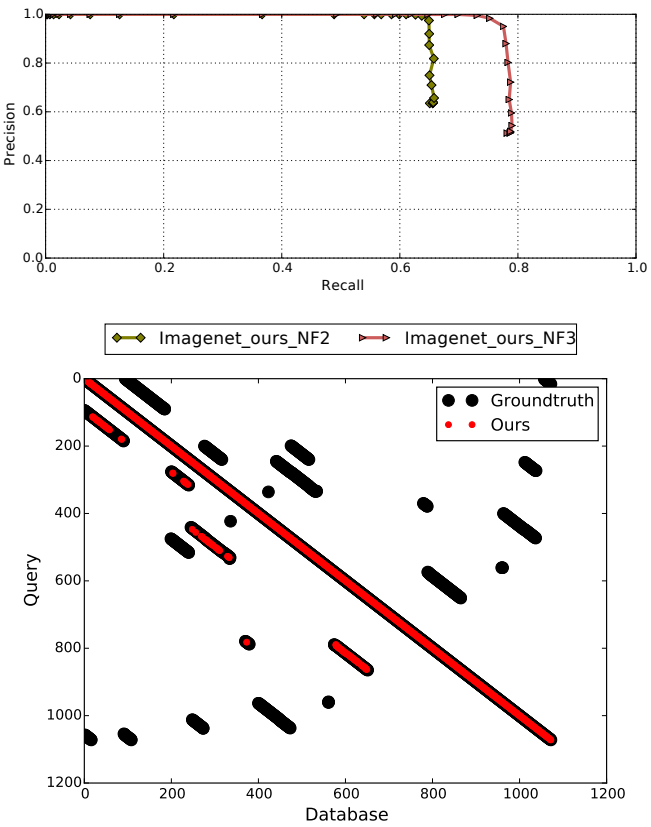

Fig. 16: *Top:* Leveraging multiple flows in our data association graph increases the recall and precision in the case of various loops in a trajectory. *Bottom:* It highlights the advantage of using multiple flows for various path hypothesis, involving intra-season loop closures. The first flow retrieves the diagonal of the trajectory where as adding more flows retrieve true matches from the loop closures which increases the recall.

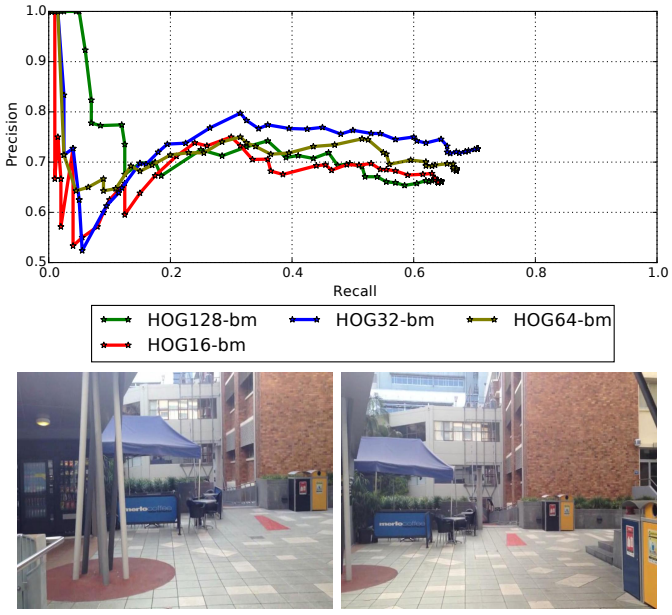

Fig. 17: *Top:* It shows the effect of cell size to compute HOG descriptors on image matching subjected to viewpoint variations. The cell size of 32×32 used in our approach outperforms smaller and larger cell sizes. *Bottom:* This figure shows the image pair successfully matched with HOG-32 and without the sequential information.

does not retrieve meaningful matches, whereas our approach retrieves 60% matches with 93% precision for the chosen threshold value as shown in Fig. 20. This is due to the keypoint-based feature descriptors used for feature matching in FABMAP2. As explained earlier, those may be suboptimal for matching across seasons, e.g., see Fig. 2. Please note that FABMAP2 uses single image for matching places whereas our method uses sequential information together with robust image features.

J. Evaluation of the Influence of Different Parameters

The next evaluation is designed to illustrate the change in performance when varying different parameters used in our approach. We start with discussing the impact of different patch sizes for HOG descriptor on the localization accuracy of the approach. We evaluated patch sizes of 32×32, 64×64, and 128×128. We discuss the effect of varying the value of K in the data association graph on the accuracy of our approach. We provide quantitative evaluations for both these parameters.

HOG Cell Size: In this experiment, we evaluate the effect of the cell size for computing HOG descriptors on the place recognition performance. The *GardenspointWalking* dataset has images captured with a hand held camera from *left* and *right* orientations to exhibit viewpoint differences as shown in Fig. 17. We evaluated four cell sizes with the side length of $(16, 32, 64, 128)$ in this experiment. Fig. 17 shows that the cell size of 32×32 outperforms the other configurations. The possible explanation for this is that the smaller cell size does not provide enough freedom for viewpoint variations, whereas the descriptions with the higher cell sizes are not discriminative enough and lead to inferior performance. This supports our choice of the cell size with the side length of 32 for a discriminative image description. The dimension of the descriptor $\mathbf{D}$ depends on the cell size $\mathcal{C}$ as,

$$\mathbf{D} = \frac{\mathcal{W}}{\mathcal{C}} \times \frac{\mathcal{H}}{\mathcal{C}} \times \mathcal{B} \tag{12}$$

where $\mathcal{W}, \mathcal{H}$ is the width and the height of the image correspondingly, and $\mathcal{B} = 128$ is the number of $\mathcal{B}$ins for the orientation of a particular cell. This configuration leads to the descriptor size of 98304 in our case.

Reachability of the Data Association Graph: The number of outgoing edges from a node in our data association graph (*parameter* K) affects the reachability of nodes from the current query to the next one. Thus, we evaluate the effect of this parameter on the localization accuracy of our approach and provide an intuition about our choice. In this experiment, we consider the same sequence as in IV-I. This sequence contains a split trajectory (unreachable with a single flow) as shown in Fig. 20. We evaluated our algorithm's performance by varying the fanout for each node between $2 \leq K \leq 6$, furthermore, we show the effect of including horizontal edges introduced in this paper and multiple flows to cope with such split trajectories. Fig. 18 shows that we gain performance boost with the fanout higher than 2, where $K = 2$ implies

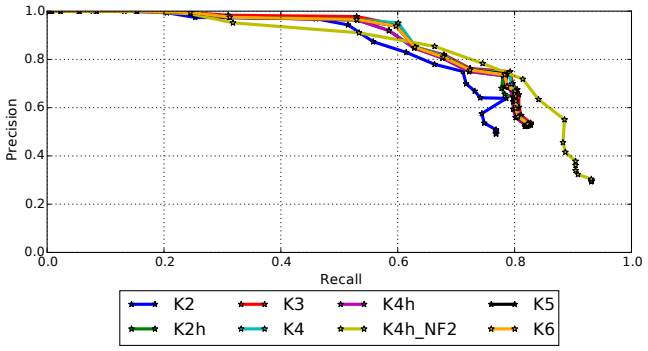

Fig. 18: This figure shows the impact of increasing the fanout (K) of each node in our data association graph. Higher values of ($K \leq 4$) provide more flexible solutions to non-linear trajectories and resulted in better performance. Increasing K>4, did not result in further gain in the accuracy. Adding horizontal edges in the graph increases the accuracy for lower values of K. Kxh implies fanout x with horizontal links and NFy implies number of flows (y) in this case.

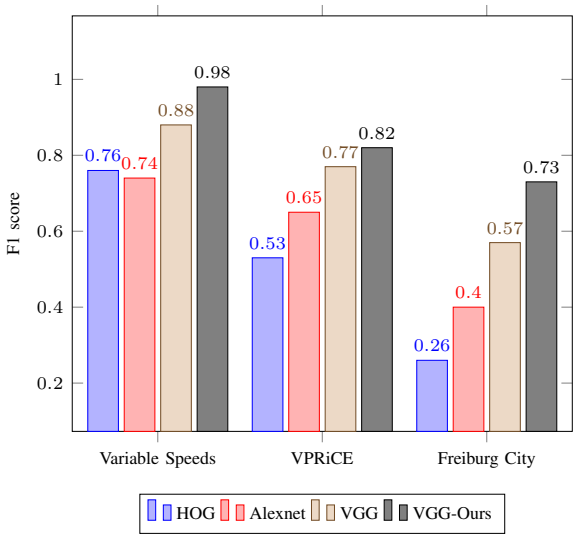

Fig. 19: This figure signifies the importance of improved convolutional architectures for robust image descriptions. VGG outperforms Alexnet architecture and handcrafted HOG features in all the cases. We achieve further gain in the performance by leveraging our graph-based sequential information.

a linear trajectory with query stoppages. We achieve the best F1 score of 0.73 and 0.77 with $K = 2$ and $K = 4$ respectively. Increasing the fanout provides more flexibility for matching non-linear trajectories. With $K > 4$, we do not gain any performance boost in this case, the performance rather drops a little for higher values supporting the choice of our fanout parameter. Furthermore, we evaluate the advantage of including horizontal links between nodes, the best F1-score with $K = 2$ and horizontal links increased from 0.73 to 0.76.

The whole set of query images cannot be matched to the database using a single flow in the localization run. Therefore, we evaluate the performance using two flows in this experiment. Fig. 18 shows that we retrieve higher number of matches at the cost of more false positives. Hence, our maximum recall increases from 81% to 93% but the overall best F1-score does not show an improvement in this case. Thus, we can state that increasing the fanout in our data association graph provides more flexibility, multiple flows enable to retrieve more matches, and the horizontal edges can improve the performance for lower values of outgoing edges.

K. Comparing Handcrafted and Deep Features

The next experiment is designed to illustrate the effect of deep features. We discuss the significance of image feature description for robust localization. We have shown that features extracted from Alexnet outperformed the handcrafted features except in the experiment IV-B. There has been a great deal of research in improving these architectures for improved image recognition and classification tasks. Here, we show that improved feature architectures can lead to further performance gains and illustrate that using the localization results on three datasets (Variable Speeds, VPRiCE, Freiburg City). Simonyan and Zisserman [46] introduce a novel deep architecture (VGG) that has shown state-of-the-art performance for image classification tasks outperforming Alexnet. Therefore, we evaluate VGG on the experiment in IV-B to emphasize that features extracted from these convolutional neural networks are more robust than handcrafted features and can lead to great advancements in place recognition and visual localization under adverse environmental conditions. Second, evaluation of VGG quantifies its performance on the VPRiCE dataset, as it consists of multiple scenarios like GardensPointWalking, Nordland and other sub-sequences. Therefore this experiment signifies its performance on most of the datasets discussed in this paper. Furthermore, we also evaluate VGG on the full city-scale dataset to provide a fair comparison under almost all the scenarios discussed in this paper. The quantified results based on the best F1-scores are shown in Fig. 19. The features from VGG provides an average of 14.3% performance boost compared to Alexnet. The localization accuracy is further improved by leveraging our graph-based sequential information by 10.3% on average compared to image-matching only based on VGG-based feature representation.

L. Runtime Evaluation

The next evaluation is designed to show the runtime requirements of our approach. Feature matching and feature extraction are generally the most time consuming operations and thus we provide only these timings here. We implemented feature matching on a GPU to cope with large databases effectively. We report timings for feature extraction of a single image and feature matching of single image-pair in Tab. I. For the Freiburg City dataset, it takes 120 ms to match a query image to the entire database of 30,790 images. Thus, our feature extraction and matching runs at 7.5 hz for the largest dataset discussed in our experiments.

To summarize, we designed various experiments to show that our approach provides robust localization under large perceptual changes due to seasonal variations, different times of the day, and in different outdoor environment types. It performs well in large scale outdoor spaces. Our method handles natural driving maneuvers such as multiple revisits

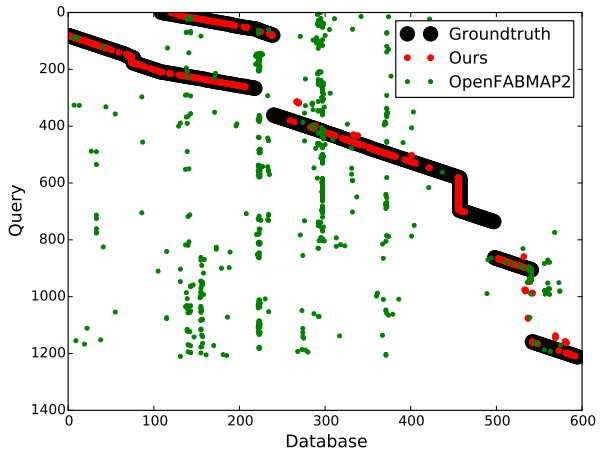

Fig. 20: We used the OpenFABMAP2 implementation in the localization -only mode i.e without creating new places. We match the query images with the database images with the default parameter settings of the implementation. Only few correct matches are found by OpenFABMAP2. This is due to the keypoint-based feature descriptors like SIFT and SURF which are not repeatable over seasons. Whereas our approach retrieves most of the matches with high precision.

Operation	GPU (ms)	CPU (ms)
Feature Extraction(DCNN)	**8.0**	31.0
Feature Extraction (HOG)	-	20.0
Feature Matching (HOG & DCNN)	**0.004**	0.1

TABLE I: GPU provides a speedup of factor 25 in feature matching and performs approximately 4 times faster than CPU-based implementation for feature extraction.

of the same place, visits to unseen areas and different driving speeds of the vehicle.

V. Conclusions

In this paper, we addressed the problem of visual localization using image sequences. We proposed a novel approach that is designed to perform localization even under substantial seasonal changes, e.g., summer vs. winter. We evaluated semi-dense image descriptions using HOG and global features from deep CNNs combined with a directed acyclic data association graph. We formulated the problem of matching image sequences over seasons as a minimum cost network flow problem and also solved the issue of dealing with non-matching image sequences that may result from collecting data at new places. Our experimental results suggest that our approach allows for accurate and robust matching across seasons and that it outperforms existing methods based on keypoint-based descriptors, descriptors from deep CNNs and also methods like OpenSeqSLAM, which also use sequence information with intensity-based image matchings.

Acknowledgments

The authors would like to thank Luciano Spinello for his valuable guidance and contributions to this work.

References

[1] M. Agrawal and K. Konolige. Frameslam: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5), 2008.

[2] R. Ahuja, T. Magnanti, and J. Orlin. *Network flows: theory, algorithms, and applications*. Prentice hall, 1993.

[3] H. Badino, D. Huber, and T. Kanade. Real-time topometric localization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2012.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

[5] H. Ben Shitrit, J. Berclaz, F. Fleuret, and P. Fua. Multi-commodity network flow for tracking multiple people. *IEEE Trans. Pattern Anal. Mach. Intell.*, 99:1, 2013.

[6] M. Bennewitz, C. Stachniss, W. Burgard, and S. Behnke. Metric localization with scale-invariant visual features using a single perspective camera. In *European Robotics Symposium*, pages 143–157, 2006.

[7] P. Biber and T. Duckett. Dynamic maps for long-term operation of mobile service robots. In *Proc. of Robotics: Science and Systems*, pages 17–24. The MIT Press, 2005.

[8] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. *Proc. of the European Conf. on Computer Vision*, pages 778–792, 2010.

[9] N. Carlevaris-Bianco and R. Eustice. Learning visual feature descriptors for dynamic lighting conditions. In *Int. Conf. on Intelligent Robots and Systems*, pages 2769–2776. IEEE, 2014.

[10] W. Churchill and P. Newman. Practice makes perfect? managing and leveraging visual experiences for lifelong navigation. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2012.

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

[12] M. Cummins and P. Newman. Appearance-based place recognition and mapping using a learned visual vocabulary model. In *Int. Conf. on Machine Learning*, 2010.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2005.

[14] A. Davison, I. Reid, N. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:2007, 2007.

[15] F. Diego, D. Ponsa, J. Serrat, and A. M. López. Video alignment for change detection. *IEEE Transactions on Image Processing*, 20(7):1858–1869, 2011.

[16] F. Diego, J. Serrat, and A. M. López. Joint spatio-temporal alignment of sequences. *IEEE Transactions on Multimedia*, 15(6):1377–1387, 2013.

[17] M. Dymczyk, S. Lynen, T. Cieslewski, M. Bosse, R. Siegwart, and P. Furgale. The gist of maps-summarizing experience for lifelong localization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, pages 2767–2773. IEEE, 2015.

[18] D. Galvez-Lopez and J. Tardos. Real-time loop detection with bags of binary words. In *Int. Conf. on Intelligent Robots and Systems*, sept. 2011.

[19] E. Garcia-Fidalgo and A. Ortiz. Vision-based topological mapping and localization methods: A survey. *Robotics & Autonomous Systems*, 64: 1–20, 2015.

[20] A. J. Glover, W.P. Maddern, M. Warren, S. Reid, M. Milford, and G. Wyeth. Openfabmap: An open source toolbox for appearance-based loop closure detection. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2012.

[21] A.J. Glover, W.P. Maddern, M. Milford, and G.F. Wyeth. FAB-MAP + RatSLAM: Appearance-based slam for multiple times of day. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, pages 3507–3512, 2010.

[22] A. Goldberg and R. Kennedy. An efficient cost scaling algorithm for the assignment problem. *Mathematical Programming*, 71(2):153–177, 1995.

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[24] E. Johns and G.-Z. Yang. Feature co-occurrence maps: Appearance-based localisation throughout the day. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2013.

[25] T. Krajník, P. deCristoforis, M. Nitsche, K. Kusumam, and T. Duckett. Image features and seasons revisited. In *Proc. of the IEEE European Conference on Mobile Robotics*. IEEE, 2015.

[26] T. Krajnik, M. Kulich, L. Mudrova, R. Ambrus, and T. Duckett. Where's waldo at time t? using spatio-temporal models for mobile robot search. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2015.

[27] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012.

[28] M. Liu and R. Siegwart. Topological mapping and scene recognition with lightweight color descriptors for an omnidirectional camera. *IEEE Transactions on Robotics*, 30(2):310–324, 2014.

[29] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004. ISSN 0920-5691.

[30] W. Maddern, A. Stewart, C. McManus, B. Upcroft, W. Churchill, and P. Newman. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2014.

[31] C. McManus, W. Churchill, W. Maddern, A. Stewart, and P. Newman. Shady dealings: Robust, long- term visual localisation using illumination invariance. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, Hong Kong, China, May 2014.

[32] C. McManus, B. Upcroft, and P. Newman. Learning place-dependant features for long-term vision-based localisation. *Robotics & Autonomous Systems*, 39(3):363–387, 2015.

[33] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65:2005, 2005.

[34] M. Milford. Vision-based place recognition: how low can you go? *Int. J. of Robotics Research*, 32(7):766–789, 2013.

[35] M. Milford and G. Wyeth. Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2012.

[36] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss. Robust visual robot localization across seasons using network flows. In *Proc. of the National Conference on Artificial Intelligence*, 2014.

[37] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard. Robust visual slam across seasons. In *Int. Conf. on Intelligent Robots and Systems*, 2015.

[38] T. Naseer, B. Suger, M. Ruhnke, and W. Burgard. Vision-based markov localization across large perceptual changes. In *Proc. of the IEEE European Conference on Mobile Robotics*, 2015.

[39] P. Neubert, N. Sunderhauf, and P. Protzel. Appearance change prediction for long-term navigation across seasons. In *Proc. of the IEEE European Conference on Mobile Robotics*, 2013.

[40] M. Paton, K. MacTavish, C. Ostafew, and T. Barfoot. It's not easy seeing green: Lighting-resistant stereo visual teach & repeat using color-constant images. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*. IEEE, 2015.

[41] A. Ranganathan. Pliss: Detecting and labeling places using online change-point detection. In *Proc. of Robotics: Science and Systems*, 2010.

[42] A. Ranganathan, S. Matsumoto, and D. Ilstrup. Towards illumination invariance for visual localization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*. IEEE, 2013.

[43] A. Ravichandran and R. Vidal. Video registration using dynamic textures. In *Proc. of the European Conf. on Computer Vision*, 2008.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.

[45] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[46] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[47] M. Smith, I. Baldwin, W. Churchill, R. Paul, and P. Newman. The new college vision and laser data set. *Int. J. of Robotics Research*, 28(5): 595–599, 2009.

[48] C. Stachniss and W. Burgard. Mobile robot mapping and localization in non-static environments. In *Proc. of the National Conference on Artificial Intelligence*, pages 1324–1329, 2005.

[49] N. Sünderhauf, F. Dayoub, S. Shirazi, B. Upcroft, and M. Milford. On the performance of convnet features for place recognition. *arXiv preprint arXiv:1501.04158*, 2015.

[50] N. Sünderhauf, S. Shirazi, A. Jacobson, E. Pepperell, F. Dayoub, B. Upcroft, and M. Milford. Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. In *Proc. of Robotics: Science and Systems*, 2015.

[51] C. Valgren and A.J. Lilienthal. SIFT, SURF & Seasons: Appearance-based long-term localization in outdoor environments. *Robotics & Autonomous Systems*, 85(2):149–156, 2010.

[52] O. Vysotska and C. Stachniss. Lazy sequences matching under substantial appearance changes. In *Workshop on Visual Place Recognition in Changing Environments at the IEEE Proceedings of the IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2015.

[53] O. Vysotska and C. Stachniss. Lazy data association for image sequences matching under substantial appearance changes. *IEEE Robotics and Automation Letters (RA-L)*, 1, 2016.

[54] O. Vysotska and C. Stachniss. Relocalization under substantial appearance changes using hashing. In *Proc. of the Workshop on Planning, Perception and Navigation for Intelligent Vehicles at the Int. Conf. on Intelligent Robots and Systems*, 2017.

[55] O. Vysotska, T. Naseer, L. Spinello, W. Burgard, and C. Stachniss. Efficient and effective matching of image sequences under substantial appearance changes exploiting gps priors. In *Proc. of the IEEE Int. Conf. on Robotics and Automation*, 2015.

[56] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2008.

Tayyab Naseer is a doctoral candidate at the University of Freiburg, Germany, working in the Lab for Autonomous Systems headed by Wolfram Burgard. He studied communication systems at Technical University of Munich and received his Masters degree in 2012. Since April 2013, he has been working towards his PhD. degree and his general research interests are visual SLAM, long-term localization, machine learning and computer vision for outdoor robots. He received the Best PhD Student award at the International Computer Vision Summer School (ICVSS) in 2016.

Wolfram Burgard is a professor for computer science at the University of Freiburg, Germany where he heads the Laboratory for Autonomous Intelligent Systems. He received his Ph.D. degree in computer science from the University of Bonn in 1991. His areas of interest lie in artificial intelligence and mobile robots. In the past, Wolfram Burgard and his group developed several innovative probabilistic techniques for robot navigation and control. They cover different aspects such as localization, map-building, path-planning, and exploration. For his work, Wolfram Burgard received several best paper awards from outstanding national and international conferences. In 2009, Wolfram Burgard received the Gottfried Wilhelm Leibniz Prize, the most prestigious German research award. In 2010 he received the Advanced Grant of the European Research Council. Wolfram Burgard is the spokesperson of the Cluster of Excellence BrainLinks-BrainTools.

Cyrill Stachniss Cyrill Stachniss is a full professor for photogrammetry and robotics at the University of Bonn. Before joining the University of Bonn, he was a lecturer at the University of Freiburg in Germany and a senior researcher at the Swiss Federal Institute of Technology. Cyrill Stachniss finished his habilitation in 2009 and received his PhD thesis from the University of Freiburg in 2006. From 2008-2013, he was an associate editor of the IEEE Transactions on Robotics, since 2010 a Microsoft Research Faculty Fellow, and received the IEEE RAS Early Career Award in 2013. Since 2015, he is a senior editor for the IEEE Robotics and Automation Letters.