IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 27, 2019, accepted March 22, 2019, date of current version April 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910150

TB-Places: A Data Set for Visual Place Recognition in Garden Environments

MARÍA LEYVA-VALLINA[1], NICOLA STRISCIUGLIO[1], MANUEL LÓPEZ-ANTEQUERA[1,2], RADIM TYLECEK[3], MICHAEL BLAICH[4], AND NICOLAI PETKOV[1]

[1]Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen, 9700 Groningen, The Netherlands
[2]MAPIR Group, Instituto de Investigación Biomédica de Málaga, University of Málaga, 29010 Málaga, Spain
[3]School of Informatics, The University of Edinburgh, Edinburgh EH8 9AB, U.K.
[4]Robotic Systems and Power Tools (CR/AER), Robert Bosch GmbH, 71272 Renningen, Germany

Corresponding author: María Leyva-Vallina (m.leyva.vallina@rug.nl)

This work was supported by the TrimBot2020 Project through the European Horizon 2020 Program under Grant 688007.

ABSTRACT Place recognition can be achieved by identifying whether a pair of images (a labeled reference image and a query image) depict the same place, regardless of appearance changes due to different viewpoints or lighting conditions. It is an important component of systems for camera localization and for loop closure detection and a widely studied problem for indoor or urban environments. Recently, the use of robots in agriculture and automatic gardening has created new challenges due to the highly repetitive appearance with prevalent green color and repetitive texture of garden-like scenes. The lack of available data recorded in gardens or plant fields makes difficult to improve localization algorithms for such environments. In this paper, we propose a new data set of garden images for testing algorithms for visual place recognition. It contains images with ground truth camera pose recorded in real gardens at different times, with varying light conditions. We also provide ground truth for all possible pairs of images, indicating whether they depict the same place or not. We also performed a thorough benchmark of several holistic (whole-image) descriptors, and provide the results on the proposed data set. We observed that existing descriptors have difficulties with scenes with repetitive textures and large changes of camera viewpoint.

INDEX TERMS Benchmark, computer vision, data set, holistic image descriptor, visual place recognition.

I. INTRODUCTION

Visual place recognition consists of recognizing a previously seen place by evaluating visual cues only from images acquired by a camera system [1], [2]. Given a query image, the most similar reference image that depicts the same scene is retrieved from a database of known place images. Subsequently, pose information related to the reference image can be used to accurately determine the position in the map where the query image was taken from.

The problem of visual place recognition can be thus formulated as distinguishing between pairs of similar and dissimilar images. It gained great interest among researchers in computer vision and became very relevant for various applications, including autonomous driving [3] and robot navigation [1], [4]–[8]. Systems for camera localization in a known environment and for loop closure detection while mapping an unknown area are also supported by visual place recognition algorithms. When a pair of images (reference and query) are recognized as depicting the same place, the relative pose between them is calculated. The reference image is expected to have a known pose in the world reference system. With this and the relative pose between the two images, the query image can be localized. Performing visual place recognition on a constrained set of visually rich and distinctive images is challenging when appearance changes are present [8] due to variations of illumination, weather conditions, camera viewpoint or when faced with textures and repetitive patterns [9].

Visual place recognition in different environments implies different challenges. Urban scenes are subject to illumination changes and partial occlusions, due to vehicles or pedestrians. In open-field scenes, challenges related to illumination or seasonal changes are present. Both urban and countryside scenes are large outdoor environments. High spatial precision for recognition or localization is not required, whereas indoor place recognition cannot allow much spatial drift, as the environment is much smaller. Furthermore, indoor

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

place recognition algorithms have to be robust to viewpoint changes to accurately recognize a room from different perspectives.

Visual place recognition in garden environments is a particularly challenging problem. Agricultural and gardening robotics are gaining increasing interest and are raising a number of challenges for computer vision algorithms [10]–[13]. For instance, algorithms are required to be robust to highly repetitive textures and viewpoint changes. In contrast to common indoor scenes, gardens have internal visual similarities, that is, one bush may look very similar to other bushes. Therefore, to successfully represent and recognize a place in a garden, a descriptor is required to ignore irrelevant parts of the image (i.e. the common background), while effectively capturing and describing important parts and their relationship within the scene (e.g. how many bushes and where they are in the scene). Despite the abundance of data sets for place and scene recognition in urban and indoor environments, which we discuss in Section II, to the best of our knowledge there are no public data sets recorded in garden-like environments that contain the mentioned challenges.

We propose a new data set for place recognition in garden environments and benchmark the performance of existing algorithms and holistic descriptors for place recognition and image recognition. The data set contains about 60k images, with ground truth provided for all pairs of images depicting the same place or not. All the images are also provided with ground truth camera pose in the garden reference system.

The paper is structured as follows. We discuss previous work and existing benchmark data sets in Section II. In Section III, we describe the proposed data set, including details on the hardware setup used for data recording, as well as on the definition of the ground truth. In Section IV, we explain the evaluation procedure, while in Section V we present and discuss the results obtained by using existing image descriptors. Finally, we draw conclusions in Section VI.

II. RELATED WORK

Various aspects and challenges of visual place recognition were studied in the recent years and several benchmark data sets were publicly released. Many of them contain urban scenes, such as KITTI [6], Dubrovnik16k and San Francisco [14], Landmark10k [15], Cambridge Landmarks [16], Tokyo Time Machine [17], Pittsburgh30k [9] and Oxford RobotCar [18]. Furthermore, Aachen [19] and Alderley [5] data sets include big changes of illumination, as they contain images recorded during day and night. The CMU Visual Localization data set [20] is composed of sub-urban environment scenes that show significant seasonal changes, that is it contains scenes depicting vegetation recorded at different times of the year. The Nordland data set [7] includes open-field scenes, and a great variance in weather and season characteristics, whereas the 7-scenes data set [21] contains images of indoor scenes.

Existing approaches to visual place recognition can be grouped in two categories: methods based on local feature descriptors and others that employ holistic scene descriptors [2]. The first group contains approaches that use and match local features (e.g. SIFT or ORB), such as the one proposed in [22] which performs place recognition for Simultaneous Localization and Mapping (SLAM), or in [23] for camera localization based on image retrieval. These methods achieved satisfactory results in indoor environments, but more recent approaches, mainly based on deep Convolutional Neural Networks (CNNs), outperformed them. The methods proposed in [4], [24] employ Bag of Visual Words representations based on the computation of histograms of the occurrence of keypoint descriptors, to perform place recognition and achieve efficient loop closure in SLAM algorithms. The main limitation of the Bag of Words representation is that it does not take into account spatial information, which might be relevant for scene recognition.

The second group contains methods based on global image descriptors, such as SeqSLAM [5], which performs place recognition by matching image sequences, instead of single images, and addressing strong illumination (day vs night) and weather (clear vs rain) variations in urban scenes. In [7], the authors expand SeqSLAM by addressing open-field images embracing a bigger environment, and with substantial seasonal changes. CNNs are able to learn global representations of images and are nowadays the leading approach in most visual recognition problems, i.e. object classification [25], [26], scene classification [27], semantic segmentation [28] and also image retrieval for place recognition. In [1], a triplet network was proposed, which takes as input two images of the same scene (i.e. positives) and one image from another scene (i.e. negative) and is trained by optimizing a triplet loss function. The network achieved high performance results with respect to illumination changes, and performed less under viewpoint changes. One of the disadvantages of CNNs is that they require a large amount of labeled examples to be effectively trained. This motivated the design of an unsupervised fine-tuning step based on Bag of Words (BoW) that was applied for selecting training images depicting similar scenes [29]. In [17] a NetVLAD architecture was proposed. Its main contribution consisted of a novel orderless pooling layer for CNN architectures inspired by the bag of words model. NetVLAD is trained using weakly-labeled triplets consisting of one reference image, a set of potential positive matches, and a set of definite negative matches. In [30], a new version of NetVLAD was proposed, which performs place recognition based on 3-D point clouds instead of 2-D images.

On a different line, end-to-end pose regression algorithms were recently proposed. Instead of relying on SLAM or image retrieval algorithms, which require to build a map of the environment, PoseNet regresses a 6DOF pose from a single image [16]. Further modifications of PoseNet were proposed, such as a Bayesian CNN for modeling uncertainty [31], or the addition of LSTMS for capturing feature correlation [32].

FIGURE 1. Four views of the TrimBot2020 garden at the Wageningen University and research campus.

These visual localization methods have been extended by using multi-task learning or combining visual localization with semantic information [33], [34].

III. DATA SET AND ITS ACQUISITION

We propose a new data set, called TB-Places (where TB stands for TrimBot), of garden images to test visual place recognition algorithms. We recorded images and ground truth pose data in the two test gardens of the TrimBot2020 project, which aims at developing the first outdoor autonomous gardening robot [13]. The gardens are located at the Wageningen University and Research (Netherlands) and in the Bosch Research Center in Renningen (Germany). The garden in Wageningen is approximately 18 $\times$ 20 meters in size and contains various elements, such as box-woods, hedges, rose bushes, trees and different terrain types, i.e. grass, woodchips and pebble stone. Some of the bushes are on a slope of 10°. The garden is surrounded by a double-line fence, to ensure safety during robot operations. In Figure 1 we show some pictures of the garden. We have recorded data in autumn 2016 and spring 2017. The garden in Renningen is about 36×20 meters in size and contains various navigable terrains (i.e. mowed grass, different types of gravel and an asphalted path). It also contains several types of obstacles, like hedges, boxwoods and concrete steps.

We publicly release images from the Wageningen garden, while use images from the Renningen garden as private test data. However, we provide a submission procedure to test place recognition algorithms on the Renningen data set.[1]

A. HARDWARE SETUP

The robot platform that we used for data recording is based on a modified Bosch Indego lawn mower robot, on which we mounted a camera rig with 360° field of view. We used an inertial measurement unit (IMU) and a TopCon laser tracker for ground truth robot and camera pose registration. The error on the accuracy of the TopCon laser is lower than 6cm, with a deviation of $\pm$1cm).

We used two configurations of the camera rig for the data recording sessions. The first configuration had eight cameras arranged in an octagon shape, as shown in Figures 2b and 2a. In both configurations, the cameras record pictures with a resolution of 752 $\times$ 480 pixels. Six cameras recorded gray-scale images while other two recorded RGB images. The second configuration of the camera rig, used for the recording session in 2017, consisted of a pentagon-shaped arrangement of five pairs of stereo cameras. A top-view of the camera rig and a sketch of the camera arrangement are shown in Figure 2c and Figure 2d, respectively. For each stereo pair, one camera recorded RGB images and the other grayscale images.

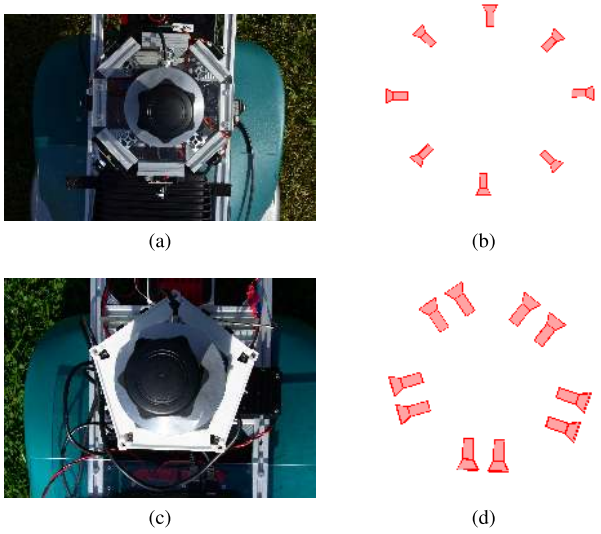


FIGURE 2. The (a) camera rig used for the data recording session in 2016 has eight cameras arranged in an (b) octagonal shape, while the (c) rig used in spring 2017 has (d) five pairs of stereo cameras in a pentagonal frame.

The acquisition of the images is synchronized by means of an FPGA, which also computes rectified and stereo images (for the camera rig with five pairs of stereo cameras) on board at 10Hz [35].

We performed an offline calibration process to compute intrinsic and extrinsic parameters of the cameras in the rig [36]. This procedure results in a chain of transformation matrices, which we use to compute the pose of each camera in the garden reference system. We represent a camera pose $\mathbf{p} = [\mathbf{t}, \mathbf{q}]$ as a translation vector $\mathbf{t} = (t^{(x)}, t^{(y)}, t^{(z)})$ and an orientation quaternion $\mathbf{q} = (q^{(x)}, q^{(y)}, q^{(z)}, q^{(w)})$.

B. TB-PLACES DATA SET

The TB-places data set that we propose contains about 60k images, organized in three sub-sets, namely the W16,

[1] https://github.com/marialeyva/TB_Places

W17 and R17 sets. The W16 set contains gray-scale images while the W17 and R17 sets contain RGB images. For experimental evaluation in this paper we consider the gray-scale version of the images in W17 and R17. As mentioned in Section III-A, each image is provided with ground truth camera pose, which indicates the position of the camera in the garden reference system when the picture has been taken.

The data set is provided with labels for all possible pairs of images in each subset, indicating whether they depict the same scene or not. In Section III-C, we provide details about the labeling process. We report detailed information about the number of images in each set, as well as the amount of similar pairs, in Table 1. We also specify the percentage of positive pairs on the total amount of pairs in each sub set.

TABLE 1. Details about the TB-Places data set, with information on each sub-set. We also specify the percentage of similar image pairs on the total number of image pairs.

Garden	Set	#imgs	#similar pairs	%similar pairs
Wageningen	W16	40752	5.12M	0.6168
Wageningen	W17	10948	330K	0.5441
Sub-total (Wageningen)		*51700*	5.44M	-
Renningen	R17	7999	150K	0.4822
Total (all)		*59699*	5.6M	-

C. GROUND TRUTH

We defined a procedure to label pairs of images whether they depict the same place or not, based on accurate measurements of the camera poses in the garden reference system. The labeling process is divided in two steps, an initial rough labeling and a label-refinement procedure.

The initial labeling step is based on two measures of pose similarity, namely a translation distance $d_t(\mathbf{t_i}, \mathbf{t_j}) = \|\mathbf{t_i} - \mathbf{t_j}\|$ and a quaternion distance $d_q(\mathbf{q_i}, \mathbf{q_j}) = 1 - (\mathbf{q_i} \cdot \mathbf{q_j})^2$, as proposed in [1]. A quaternion distance can be converted to degrees by means of the formula $\Delta\phi_{i,j} = \Phi(\mathbf{q_i}, \mathbf{q_j}) = \arccos\left(-2\, d_q(\mathbf{q_i}, \mathbf{q_j}) + 1\right) \cdot 180/\pi$.

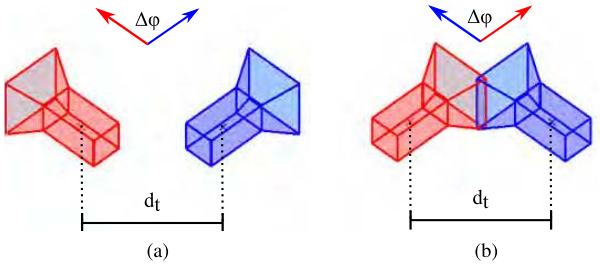


FIGURE 3. Example of ambiguous cases where translation and orientation distances are the same: the cameras are oriented at (a) different places or at (b) the same place.

These distance measures are meaningful in those cases where the camera locations are very close and orientations are very similar. In cases where there are larger camera viewpoint variations, their values can lead to ambiguous situations. In Figure 3, we show an example of an ambiguous case. Two pairs of cameras, whose pose vectors have same translation and orientation distances, are oriented at different directions. In order to resolve such cases, we take into account an approximation of the 2D fields of view of the cameras on the horizontal plane, which we represent by a radius r and the camera lens aperture angle β. This choice is motivated by the fact that the robot navigates a planar surface and the camera rig can be tilted only by few degrees. We compute the field of view area $fov^{(\mathbf{p_i})} = 1/2 \cdot \beta r^2$ associated to the camera pose $\mathbf{p_i}$ as the area of a circle sector of radius r meters and angle β radians, where the pose vector $\mathbf{p_i}$ provides information on the position $\mathbf{t_i}$ and orientation $\mathbf{q_i}$ of the field of view circle sector with respect to the reference system. For a pair of camera poses $\mathbf{p_i}$ and $\mathbf{p_j}$, we compute a measure of the percentage of overlap (FOVO) of their field of view areas (0 means no overlap, while 1 indicates a perfect overlap) as:

$$\text{FOVO}(\mathbf{p_i}, \mathbf{p_j}) = \frac{fov^{(\mathbf{p_i})} \cap fov^{(\mathbf{p_j})}}{fov^{(\mathbf{p_i})} \cup fov^{(\mathbf{p_j})}} \tag{1}$$

An example of the computation of this measure is illustrated in Figure 4, where the darker region indicates the region where the fields of view overlap. We use the FOVO measure to refine the labeling process, as explained in the following.

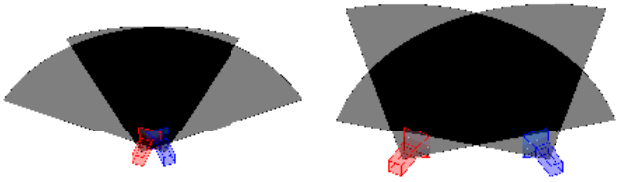
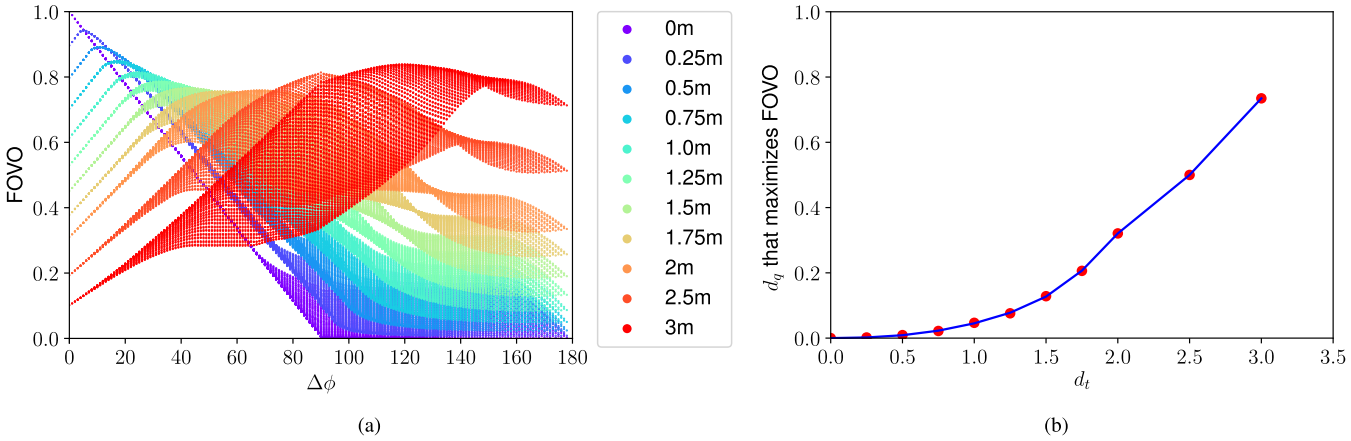


FIGURE 4. Examples of two field of view overlaps with (a) $d_t = 0.2$, $\Delta\phi = 40°$, *FOVO* = 0.71 and (b) $d_t = 1.0$, $\Delta\phi = 60°$, *FOVO* = 0.75.

For a pair of images with a small translation distance d_t, a small quaternion distance d_q is also required so that the FOVO measure is large (i.e. close to 1). If the translation distance is large, the rotation distance can be larger as well, so that the two camera poses share a bigger part of the field of view. We considered different translation distances and their possible combinations with several rotation distances in order to find a satisfying FOVO measure. We refer at Figure 5a where we show the relation between the considered distance measures. We considered several values for translation and quaternion distances, and calculated the possible FOVO measures. We computed a function $f(\cdot)$ by a polynomial interpolation (Fig. 5b) that computes, for a given value of the translational distance between two poses, the quaternion distance that would guarantee the maximum FOVO measure. Given a translation distance, the function f computes the rotation distance that maximizes the FOVO measure $\hat{d}_q^{(i,j)} = f(d_t(\mathbf{t_i}, \mathbf{t_j}))$.



(a) (b)

FIGURE 5. Relationship between the (a) camera viewpoint angle and FOV overlap for different translation distance values. (b) Approximation of function f by a polynomial regression, that computes the quaternion distance that maximizes the FOVO for a given translation distance.

Finally, we defined that a pair of images with associated camera poses $\mathbf{p_i} = (\mathbf{t_i}, \mathbf{q_i})$ and $\mathbf{p_j} = (\mathbf{t_j}, \mathbf{q_j})$ are depicting the same place if they fulfill the following conditions:

1) $d_t(\mathbf{t_i}, \mathbf{t_j}) \leq 1.5m$
2) $|\Phi\left(d_q(\mathbf{q_i}, \mathbf{q_j})\right) - \hat{d}_q^{(i,j)}| < 15°$
3) $\text{FOVO}(\mathbf{p_i}, \mathbf{p_j}) \geq 0.53$

We consider two images to depict the same place if their ground truth camera poses 1) are close enough, 2) have a relative rotation angle that guarantees an overlap of at least 53 % of the area of their fields of view. The values were determined by experimental observations, in which we took into account the presence and arrangement of close objects, which are relevant to recognize places in such environments. Some examples can be seen in Figure 6. We provide the ground truth in the form of a binary compressed matrix, where entries with value equal to zero indicate a negative pair (that is, a pair of images not depicting the same scene), and a one indicates a positive match, that is, two images showing the same place. In Figure 6, we show some examples of images from the TB-Places data set, consisting in three reference images (leftmost images), two positive (second and third column) and one negative (rightmost column) images.

IV. EVALUATION

In the following, we provide details about the holistic image descriptors that we employ in our comparative analysis, the experiments that we designed and the metrics that we employed to evaluate the results on the proposed TB-Places data set.

A. IMAGE DESCRIPTORS

We consider the holistic image descriptor proposed in [1], where the authors compute a 128-D feature vector with a lightweight deep neural network, called DepredNet, trained for visual place recognition in outdoor environments. We also consider Histogram of Oriented Gradients (HOG), with feature vectors of size 19200 [37]. We include in the evaluation the descriptors computed as output of several layers of the VGG and AlexNet CNNs, pre-trained on the Places365 data set [27]. Furthermore, we consider a descriptor computed by using the NetVLAD-VGG16 CNNs with whitening, pre-trained on the Pittsburgh30K and Tokyo Time Machine data sets [17]. The size of the feature vector computed by NetVLAD is 4096 elements. We report a summary of the employed descriptors and their feature vector size in Table 2.

TABLE 2. Details on the whole-image descriptors used for the baseline performance comparison analysis.

Descriptor	Size	Pre-trained on
DepredNet [1]	128	KITTI [6] and Nordland [7]
HOG [37]	19200	-
VGG16, pool4 [26]	100352	Places365 [27]
VGG16, pool5 [26]	25088	Places365
AlexNet, pool2 [25]	43264	Places365
AlexNet, pool5 [25]	9216	Places365
NetVLAD [17]	4096	Pittsburgh30k [9]
NetVLAD [17]	4096	TokyoTM [17]

B. EXPERIMENTS

For the baseline experimental analysis on the proposed TB-Places data set, we defined the following tests:

- **Training set: W16, Test set: W17**. This experiment is meant to evaluate robustness of place recognition algorithms against different illumination conditions and variations of the garden due to seasonal changes. We also evaluate robustness to camera viewpoint variation.
- **Training set: W16, Test set: R17**. This experiment aims to evaluate the generalization capability of place recognition descriptors to new garden environments. Since training is usually a time consuming and computationally expensive procedure and requires large amount of ground truth data, it is very desirable that

FIGURE 6. Example images from the TB-Places data set. For each reference image (left column), we show two positive matches (second and third column, surrounded by a green solid line) and one negative image (surrounded by a red dashed line). In (a), the translation distances between the reference and the query images are, from left to right, $d_t = (0.99, 0.77, 12.47)m$, the quaternion distances are $d_q = (0.0387, 0.0006, 0.0050)$, which correspond to $\Delta\phi = (22.68°, 2.72°, 8.10°)$, and the FOVOs are (0.7820, 0.7289, 0.1941). In (b), the translation distances between the reference and the query images are $d_t = (0.70, 0.78, 7.01m)$, the quaternion distances $d_q = (0.0301, 0.0028, 0.0442)$, which correspond to $\Delta\phi = (19.98°, 6.02°, 24.28°)$, and the FOVOs are FOVO $= (0.5491, 0.5789, 0)$. In (c) the translation distances between the reference and the query images are $d_t = (0.43, 0.68, 9.21)m$, the quaternion distances $d_q = (0.0007, 0.0026, 0.1146)$, which correspond to $\Delta\phi = (3.04°, 5.81°, 39.57°)$, and the FOVOs are (0.7545, 0.8048, 0.0004). In (d) the translation distances between the reference and the query images are $d_t = (0.76, 0.78, 0.90)m$, the quaternion distances $d_q = (0.0506, 0.0212, 0.0187)$, which correspond to $\Delta\phi = (26.01°, 16.73°, 15.71°)$, and the FOVOs are (0.5353, 0.6312, 0.5130).

place recognition algorithms and descriptors generalize robustly to different environments.

We performed different experiments: both in the training and test phases, we extracted descriptors from the images in the TB-Places data set and normalized them using one of the following feature vector normalization: none, L_1 and L_2. Furthermore, we explored the use of Principal Component Analysis (PCA) on the descriptors computed on the training set. We computed the principal components and considered a number of dimensions that contain the 95% of the variance of the training data. The purpose of using the PCA in this analysis is twofold. We aim at 1) showing that the training image data contains useful information to train models that are meaningful for place recognition in gardens and 2) evaluating if features computed by pre-trained networks for place recognition generalize to and are robust for the analysis of garden scenes. For each feature normalization approach, we performed experiments with and without PCA. In Table 3, we report the size of the considered image descriptors before and after applying PCA on the training data. One can observe that, for all the considered descriptors, most of the features are not useful for effective descriptions of garden images. The PCA analysis, indeed, significantly reduced the number of dimensions that contain relevant information for the task at hand (between 37.5% and 98.45% of the feature dimensions, according to the specific descriptor and feature normalization employed).

In the evaluation phase, we projected the feature vectors extracted from the test images onto the PCA space computed on the training data. Subsequently, we computed all

TABLE 3. Dimension reduction performed by computing principal component analysis (PCA) on the training data of the TB-Places data set. We evaluated different feature normalization, namely no normalization, L1 and L2 norms. In the last column, we report the percentage of dimensions that are discarded by the PCA.

Descriptor	Norm	Size		Dimension reduction (%)
		No PCA	PCA	
DepredNet	-	128	80	37.50
	L_1		73	42.97
	L_2		77	39.84
HOG	-	19200	455	97.63
	L_1		533	97.22
	L_2		297	98.45
VGG16, pool4	-	100352	11668	88.37
	L_1		6621	93.40
	L_2		25648	74.44
VGG16, pool5	-	25088	3758	85.02
	L_1		2237	91.08
	L_2		11395	54.58
AlexNet, pool2	-	43264	4822	88.85
	L_1		406	99.06
	L_2		9613	77.78
AlexNet, pool5	-	9216	1914	79.23
	L_1		639	93.07
	L_2		3788	58.90
NetVLAD (Pittsburgh30k)	-	4096	2459	39.97
NetVLAD (TokyoTM)	-	4096	2562	37.45

the pairwise distance between descriptors. We considered the Cosine dissimilarity and Euclidean distance measures. We computed the two distance measures for each combination of descriptor and feature normalization, with and without applying the PCA transformation. If the computed distance is higher than a certain threshold t, the two images are considered to depict different places, otherwise, the images are classified as showing the same place. During the evaluation, we studied the effect of different values of the distance threshold t.

C. PERFORMANCE MEASURES

We evaluate the performance of the considered descriptors by computing the precision (P), recall (R) and F_1-score for the classification of pairs of images as depicting or not the same place:

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$
$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

where TP are true positives, FP are false positives, and FN are false negatives. A positive class example is defined in the ground truth as a pair of images that depict the same place, otherwise it is a negative sample. We also compute the precision-recall curve by varying the value of the threshold t on the distance of image descriptors (small distance indicates high image similarity). We compute the Average Precision (AP) as the mean of the Precisions obtained for each threshold weighted by the increase in Recall achieved at threshold n with respect to the one achieved at threshold $n-1$. It is defined as:

$$AP = \sum_n (\mathrm{R}_n - \mathrm{R}_{n-1}) P_n$$

V. RESULTS AND DISCUSSION

In Table 4, we report the AP that we achieved on the W17 and R17 test sets. We also report the results that we obtained when using the descriptors projected onto the PCA space computed on the W16 training set. We measured very low recognition performance achieved by DepredNet (W17: $AP = 4.30\%$; R17: 5.09%) and HOG descriptors (W17: $AP = 9.31\%$; R17: 3.87%). Internal representations computed in popular CNN architectures, such as AlexNet and VGGNet, do not perform well to garden scenes and the task of visual place recognition too.

We obtained the best results by using the NetVLAD CNN architecture pre-trained on the Pittsburgh30k data set. NetVLAD is specifically designed for place recognition and to be robust against camera viewpoint changes. We achieved 21.60 % average precision on the W17 test set, and 15.68 % on the R17 test set. These results are also due to the very unbalanced data sets, where positive image pairs are much less represented than the negative ones. When applying the PCA transformation learned from the W16 training set data to the descriptors computed on the test images, the performance results increase both on the W17 ($AP = 23.61$ %) and the R17 ($AP = 16.86$ %) test sets.

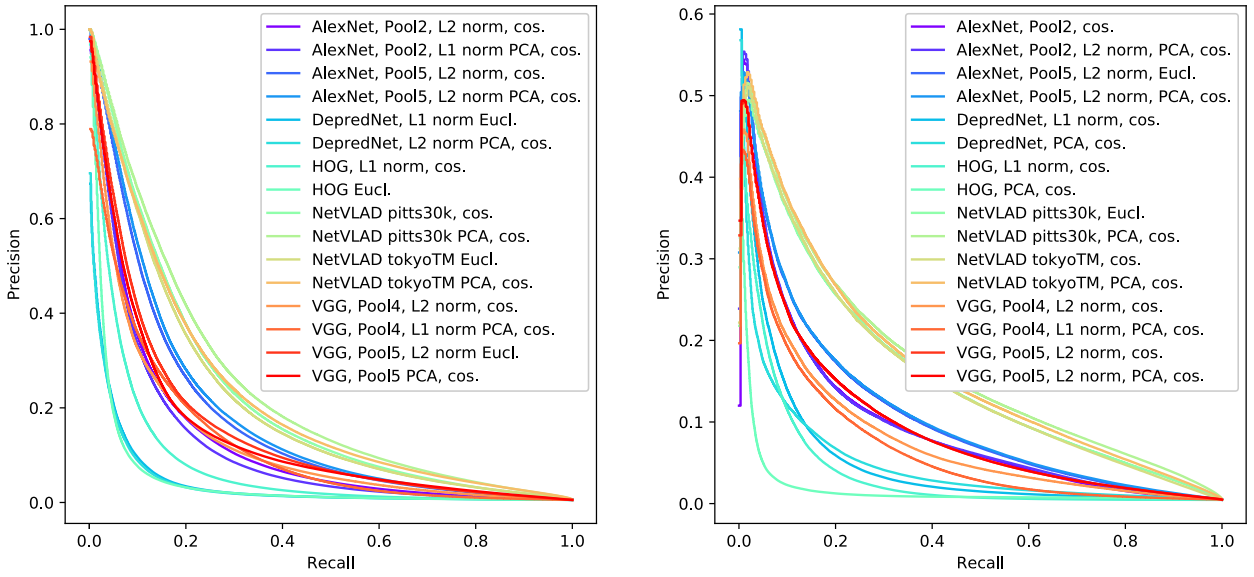
In general, we observed that the use of the cosine dissimilarity contributes to slightly better results than using Euclidean distance, and that feature vector normalization also improves the performance results. In most of the cases the Average Precision increases after applying the PCA projection learned from the training data. This confirms the usefulness of training data, which can be used to learn image recognition models specific for garden data and achieve better results.

In Figure 7a and Figure 7b, we show the precision-recall curves that we obtained on the W17 and R17 test sets, respectively. The average scarce performance of various descriptors are an indication that garden scenes have very peculiar characteristics, which existing methods or models are not robust to. Non-sharp object boundaries and the abundance of texture and green color are hard challenges for existing algorithms for holistic image description and visual place recognition. In order to support visual navigation and localization of robots in such environments, further studies and improvement of the existing approaches on the challenges provided by garden scenes are required.

TABLE 4. Average precision that we achieved using the considered descriptors, with and without normalization and PCA, using Cosine (C), and Euclidean (E) distances, on the Wageningen 2017 and Renningen 2017 test sets.

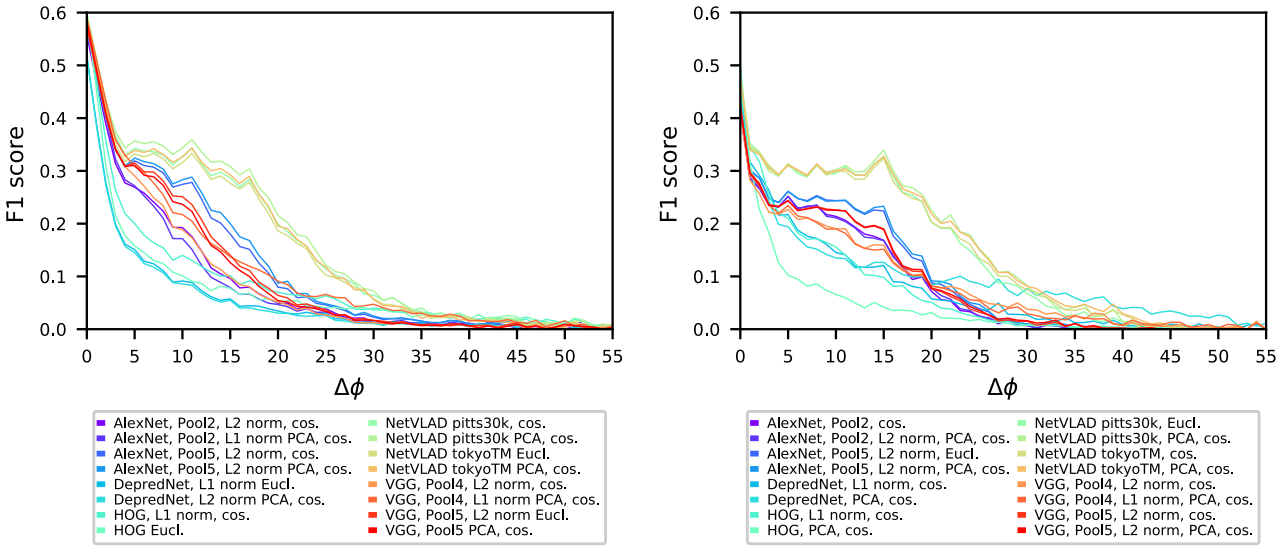
Descriptor	Norm	Distance	W17		R17	
			No PCA	PCA	No PCA	PCA
Deprednet	-	C	0.0407	0.0430	0.0509	0.0489
	-	E	0.0391	0.0396	0.0474	0.0492
	L_1	C	0.0407	0.0400	0.0509	0.0520
	L_1	E	0.0411	0.0417	0.0506	0.0505
	L_2	C	0.0407	0.0396	0.0509	0.0515
	L_2	E	0.0407	0.0413	0.0509	0.0505
HOG	-	C	0.0772	0.0415	0.0191	0.0176
	-	E	0.0388	0.0452	0.0191	0.0200
	L_1	C	0.0772	0.0846	0.0384	0.0387
	L_1	E	0.0660	0.0931	0.0244	0.0296
	L_2	C	0.0455	0.0071	0.0293	0.0083
	L_2	E	0.0071	0.0069	0.0083	0.0079
AlexNetPool2	-	C	0.1135	0.1240	0.0938	0.1000
	-	E	0.0909	0.0929	0.0921	0.0905
	L_1	C	0.0978	0.0960	0.0705	0.0674
	L_1	E	0.0193	0.0203	0.0280	0.0301
	L_2	C	0.1321	0.1486	0.0921	0.0975
	L_2	E	0.0548	0.0550	0.0644	0.0048
AlexNetPool5	-	C	0.1685	0.1848	0.1081	0.1187
	-	E	0.1380	0.1512	0.1079	0.1160
	L_1	C	0.1685	0.1482	0.1081	0.0960
	L_1	E	0.1654	0.1553	0.0983	0.0878
	L_2	C	0.1685	0.1771	0.1081	0.1098
	L_2	E	0.1685	0.1757	0.1081	0.1096
VGGPool4	-	C	0.1090	0.1115	0.0758	0.0756
	-	E	0.0511	0.0512	0.0499	0.0495
	L_1	C	0.1216	0.1206	0.0716	0.0692
	L_1	E	0.0144	0.0144	0.0178	0.0178
	L_2	C	0.1230	0.1299	0.0799	0.0815
	L_2	E	0.0277	0.0277	0.0309	0.0310
VGGPool5	-	C	0.1473	0.1378	0.0951	0.0980
	-	E	0.0761	0.0766	0.0723	0.0720
	L_1	C	0.1473	0.1494	0.0951	0.0936
	L_1	E	0.1315	0.1298	0.0891	0.0885
	L_2	C	0.1473	0.1467	0.0951	0.0952
	L_2	E	0.1473	0.1472	0.0951	0.0948
NetVLAD (Pittsburgh30k)	-	C	**0.2160**	**0.2361**	0.1568	**0.1686**
	-	E	**0.2160**	0.2114	0.1568	0.1556
NetVLAD (TokyoTM)	-	C	0.2070	0.2190	**0.1570**	0.1651
	-	E	0.2070	0.2041	**0.1570**	0.1545

We also evaluated the robustness of the considered image descriptors with respect to variations of the camera viewpoint. For a given descriptor and feature normalization, we select the value t^* of the threshold t that contributes to the highest value of the F_1-score. We thus analyzed the performance results achieved by using the considered descriptors as the viewpoint angle between the positions of the cameras where the pictures were taken from increases. We compute the viewpoint angle by converting the quaternion distance d_q to an angle on the horizontal plane and measuring its degrees. We indicate with $\Delta\phi$, the viewpoint angle difference in degrees. We show the results of this evaluation, in Figure 8. We report the F1-score achieved by the considered image descriptors for increasing values of the viewpoint angle difference $\Delta\phi$. We observe a general decrease of performance with increasing viewpoint angle between cameras. In most cases, the recognition



(a) (b)

FIGURE 7. Precision recall curves achieved on the (a) W17 and (b) R17 test sets.



(a) (b)

FIGURE 8. F_1-score as the camera viewpoint difference($\Delta\phi$) increases for (a) W17 and (b) R17 data sets.

performance results drops to 0 (i.e. images are not recognized as depicting the same place) when the viewpoint angle between camera $\Delta\phi$ is greater than 30°. NetVLAD is the method that shows the highest robustness with respect to camera viewpoint changes, maintaining stable recognition performance when the camera viewpoint angle is lower that 20°. Generally, the considered descriptors show reasonable performance when the viewpoint angle between cameras is very small (e.g. NetVLAD and VGG descriptors achieved an F_1-score of 0.6 on the W17 and of 0.5 on the R17 test sets). The drop of recognition performance under increasing camera viewpoint angle demonstrates that existing holistic descriptors are not robust to perspective changes in the images. This stimulates the necessity of designing models that are able to deal with such geometric changes, which sum up to the challenges of garden scenes.

VI. CONCLUSIONS

We proposed a novel data set to test algorithms for visual place recognition in garden environments, called TB-Places. The data set is composed of about 60k images recorded in real gardens, in the context of the TrimBot2020 project, and

it's designed for the evaluation of robustness of algorithms to strong camera viewpoint changes and to largely textured scenes. We provided ground truth labels for all possible image pairs, indicating whether they depict the same place or not.

We compared the performance of several holistic image descriptors on the task of place recognition under camera viewpoint changes. The baseline results that we obtained show that existing descriptors suffer from challenging conditions of garden scenes and are not robust, or are robust to a small extent, to perspective changes in the images. The decrease of performance results as the camera viewpoint angle increases, indeed, indicates a lack of robustness against this kind of scene changes.

The challenges contained in the proposed data set, such as the prevalent green-color and textured environment and the large camera viewpoint difference under which the scenes are depicted, require the design of new robust image descriptors to be embedded in systems for robot localization or navigation in gardens.

REFERENCES

[1] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, "Appearance-invariant place recognition by discriminatively training a convolutional neural network," *Pattern Recognit. Lett.*, vol. 92, pp. 89–95, Jun. 2017.

[2] S. Lowry *et al.*, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[3] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, "Shady dealings: Robust, long-term visual localisation using illumination invariance," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Jun. 2014, pp. 901–906.

[4] M. Cummins and P. Newman, "FAB-MAP: Probabilistic localization and mapping in the space of appearance," *Int. J. Robot. Res.*, vol. 27, no. 6, pp. 647–665, 2008.

[5] 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. Automat.*, May 2012, pp. 1643–1649.

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[7] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? Challenging SeqSLAM on a 3000 km journey across all four seasons ," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, Aug. 2013, pp. 1–3.

[8] T. Sattler *et al.*, "Benchmarking 6DOF outdoor visual localization in changing conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2018, pp. 8601–8610.

[9] A. Torii, J. Sivic, M. Okutomi, and T. Pajdla, "Visual place recognition with repetitive structures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2346–2359, Nov. 2015.

[10] C. W. Bac, E. J. van Henten, J. Hemming, and Y. Edan, "Harvesting robots for high-value crops: State-of-the-art review and challenges ahead," *J. Field Robot.*, vol. 31, no. 6, pp. 888–911, 2014.

[11] N. Ohi *et al.*, "Design of an autonomous precision pollination robot," in *Proc. IEEE IROS*, Oct. 2018, pp. 7711–7718.

[12] A. Walter *et al.*, "Flourish-a robotic approach for automation in crop management," in *Proc. ICPA*, Jun. 2018, pp. 24–27.

[13] N. Strisciuglio *et al.*, "Trimbot2020: An outdoor robot for automatic gardening," in *Proc. 50th Int. Symp. Robot. ISR*, Jun. 2018, pp. 1–6.

[14] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *Proc. ECCV*, Sep. 2010, pp. 791–804.

[15] Y. Li, N. Snavely, D. P. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3D point clouds," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 15–29.

[16] A. Kendall, M. Grimes, and R. Cipolla, "PoseNet: A convolutional network for real-time 6-DOF camera relocalization," in *Proc. IEEE ICCV*, Dec. 2015, pp. 2938–2946.

[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE CVPR*, Jun. 2018, pp. 1437–1451.

[18] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford Robotcar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2016.

[19] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited," in *Proc. BMVC*, 2012, vol. 1, no. 2, p. 4.

[20] H. Badino, D. Huber, and T. Kanade, "Visual topometric localization," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 794–799.

[21] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images," in *Proc. IEEE CVPR*, Jun. 2013, pp. 2930–2937.

[22] P. Newman and K. Ho, "Slam-loop closing with visually salient features," in *Proc. IEEE ICRA*, Apr. 2005, pp. 635–642.

[23] 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.

[24] M. Cummins and P. Newman, "Highly scalable appearance-only SLAM-FAB-MAP 2.0," *Robot., Sci. Syst.* vol. 5, p. 17, Jun. 2009.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, " ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Feb. 2012, pp. 1097–1105.

[26] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[27] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[28] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[29] F. Radenovic, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. ECCV*, Oct. 2016, pp. 3–20.

[30] M. A. Uy and G. H. Lee, " PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition," in *Proc. IEEE CVPR*, Jun. 2018, pp. 4470–4479.

[31] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Proc. IEEE ICRA*, May 2016, pp. 4762–4769.

[32] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *Proc. IEEE ICCV*, Oct. 2017, pp. 627–637.

[33] P. Wang, R. Yang, B. Cao, W. Xu, and Y. Lin, "DeLS-3D: Deep localization and segmentation with a 3D semantic map," in *Proc. IEEE CVPR*, Jun. 2018, pp. 5860–5869.

[34] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler, "Semantic visual localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6896–6906.

[35] D. Honegger, T. Sattler, and M. Pollefeys, "Embedded real-time multi-baseline stereo," in *Proc. IEEE ICRA*, Jun. 2017, pp. 5245–5250.

[36] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. IEEE IROS*, Nov. 2013, pp. 1280–1286.

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

MARÍA LEYVA-VALLINA received the M.Sc. degree in artificial intelligence from the Polytechnic University of Catalonia. She is currently pursuing the Ph.D. degree with the Intelligent Systems Group, Bernoulli Institute, University of Groningen. Her main research interests include representation learning and computer vision.

NICOLA STRISCIUGLIO received the Ph.D. degree *cum laude* in computer science from the University of Groningen, The Netherlands, in 2016, and the Ph.D. degree in information engineering from the University of Salerno, Italy, in 2017.

He is currently a Postdoctoral Researcher with the Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen. He has been the General Co-Chair of the First and Second International Conference on Applications of Intelligent Systems (APPIS), in 2018 and 2019. His research interests include pattern recognition, machine learning, signal processing, and computer vision.

MANUEL LÓPEZ-ANTEQUERA received the B.Sc. and M.Sc. degrees in electrical engineering from the University of Málaga, Spain, in 2009 and 2013, respectively.

He is currently a Computer Vision Engineer at Mapillary. He is currently pursuing the Ph.D. degree with the Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen, and the Machine Perception and Intelligent Robotics Group, University of Málaga. His research interest includes the intersection of machine learning and geometric computer vision for applications such as robot localization.

RADIM TYLECEK received the Ph.D. degree in artificial intelligence from Czech Technical University in Prague, in 2016.

He is currently a Postdoctoral Researcher with The University of Edinburgh, where he was working with B. Fisher. He is interested in vision for robotics including 3D reconstruction and fusion, and also regular semantic structures in images such as facades. He has organized workshops and challenges on 3D reconstruction and semantics at ICCV and ECCV.

MICHAEL BLAICH received the Ph.D. degree in computer science from the University of Oldenburg, Germany, in 2016. He is currently a Research Engineer in robotics and future systems with Robert Bosch GmbH, and also a Lecturer in robotics with the University of Applied Sciences, Konstanz, Germany. His research interests include simultaneous localization and mapping (SLAM), robot navigation, and computer vision.

NICOLAI PETKOV received the Dr.Sc. (tech.) degree in computer engineering (informationstechnik) from the Dresden University of Technology, Dresden, Germany. Since 1991, he has been a Professor of computer science and the Head of the Intelligent Systems Group, University of Groningen. He has authored two monographs. He has authored or coauthored over 150 scientific papers. He holds four patents. His current research interests include pattern recognition, machine learning, data analytics, and brain-inspired computing, with applications in healthcare, finance, surveillance, manufacturing, robotics, and animal breeding. He is an Editorial Board Member of several journals.