Deep Embeddings-based Place Recognition Robust to Motion Blur

Piotr Wozniak
Rzeszów University of Technology
Al. Powstańców Warszawy 12, 35-959 Rzeszów, Poland

Bogdan Kwolek✉
AGH University of Science and Technology
30 Mickiewicza, 30-059 Kraków, Poland
bkw@agh.edu.pl

Abstract

In this work we present an algorithm for severe (unknown) blur detection on RGB images. On salient CNN-based regional representations we calculate local features that are then fed to calibrated classifiers in order to estimate blur intensity. We perform scene classification and show that considerable gain in classification performance can be obtained owing to information on blur presence. We calculate global descriptors of the scene that are then fed to image retrieval engine that uses blur detection, scene category and minimum spanning tree to decide if current query image is relevant or irrelevant in context of place recognition. We show that information about blur and scene category improves mean average performance. We introduce a freely available challenging dataset both for blur detection and place recognition. It contains both images with severe blurs and sharp images with 6-DOF viewpoint variations, which were recorded using a humanoid robot.

1. Introduction

Visual Place Recognition (VPR) refers to capability of recalling a previously visited place using visual input, under changing viewpoint, varying illumination conditions, while requiring as less as possible computational power and memory storage [17]. In the last decade, several methods for visual place recognition have been proposed [17, 6]. Variations in viewpoint and appearance due to motion blur, shadows, occlusions and illumination changes render VPR as a challenging problem for autonomous robots. For humanoid robots, due to different poses while taking images the differences between image content can be considerable even when passing the same route. The degree of viewpoint variation that arises during scene perception by a humanoid robot is appreciably more complex than viewpoint variations experienced by mobile robots [27]. When a humanoid robot is walking, squatting or turning its head moves in a jerky and sometimes unpredictable way [29].

Motion blur, one of the major problems for feature-based SLAM systems, might cause location losses and inaccuracies during map formation. Most of benchmark data for visual place recognition include lateral or 3D variations of viewpoint. 6-DOF viewpoint variations are included in outdoor images of 24/7 Query dataset [38]. Recently, 6-DOF viewpoint change has been included in the Shopping street dataset [18] that is targeted for aerial place recognition. Having on regard long-term operation, aside of motion blur the appearance variations caused by day-night cycles can have considerable impact on place recognition performance. In this context it is worth noting that most of VPR benchmark data can be categorized as time-based, which means that frames have been acquired and stored at a fixed FPS (frames per second) rate of utilized video camera. Usually, they are acquired under assumption of non-zero speed of the moving camera. For instance, in [12] a frame is acquired and then stored every few meters to obtain a representation of a new place. One of disadvantages of both time- and distance-based systems are vast requirements for image storage and retrieval. Thus, such approaches can be fragile for long-term robot missions with day-night cycles.

The problem of qualitative robot localization consists in recognizing the place where the robot is located [15]. Ideally, during exploration of the environment the robot should learn from experience and then recognize previously observed places in known environment(s) and optionally categorize previously not visited places into new rooms. Such a task is closely related to semantic localization that aims at determining by a robot its location semantically in relation to objects or regions in the scene rather than estimating and reporting 6-DOF pose or position [4].

A typical approach to VPR consists in extracting single

frame or a few frames from the image-stream and then providing such data to an image retrieval engine. An approach proposed in [9] operates on sequences of non-overlapping images and decides if each one belongs to already visited place. To enhance topological place discovery and reduce the search space a clustering of images on the basis of place topics has been proposed in [24]. Instead of using individual images or image descriptors a correlation-based matching on entire sequences was introduced in [22]. In a recent work [11], multilevel descriptors have been used in VPR for the visually impaired people. Several handcrafted local and global feature descriptors were proposed to represent places [17]. Since seminal work on using CNNs for place recognition [7], more and more data-driven image description approaches are developed. Performance of such algorithms has been studied in [33]. Recently, in [1] a VLAD [2] layer that can be trained in end-to-end fashion, specifically for place recognition has been introduced. Considerable potential of VLAD has recently been proved in [41], where a versatile comparison of ten VPR systems revealed the NetVLAD as the best overall performing technique.

Recently, Intel developed a AI-powered backpack [20] that is breakthrough technology for visually impaired to navigate world around them. This user-friendly interactive device helps detecting common objects such as traffic signs, crosswalks, hanging obstacles, moving objects, and changing elevations, all while running on a low-power device. As the user moves through the environment, this versatile and powerful AI device audibly communicates him/her about common obstacles. User localization is a vital component of indoor blind navigation. To the best of our knowledge, there are no previous studies addressing place recognition on sequences of images acquired by body-worn cameras, including quantitative analysis on a dedicated dataset with images contaminated by unknown motion blur.

Motivated by lack of adequate dataset including images with variations arising during locomotion of walking robots and visually impaired equipped with body-worn cameras, especially comprising images with severe blurs we used a camera mounted on head of humanoid robot for recording a dataset. In order to fulfill the existing gap as well as fulfill requirements regarding deep learning, considerable efforts were devoted for manual classification of images as sharp and blurry. Moreover, query images with relevant and irrelevant images were selected to benchmark the algorithms for place recognition. In order to cope with place recognition on images with unknown blur we propose an effective algorithm for blur detection. We develop effective approach for scene classification and demonstrate its usefulness for visual place inference. We propose an algorithm for place recognition that on the basis of two most relevant images, which are selected by two best methods established dynamically, determines the final most relevant image.

2. Relevant work

Indoor scene recognition is a challenging open problem [43, 16] and range of approaches have been proposed in the last decade. The most commonly employed handcrafted global descriptor is GIST [26]. With the development and advancement of deep learning there has been a major paradigm shift consisting in focusing on neural network activations-based descriptors for scene as well as place recognition. As demonstrated in [39], features extracted from CNN layers and used as global descriptors hold considerable potential. Geometric features like vertical lines can also be very valuable representations of buildings or objects like doors in outdoor/indoor environments [3].

Recent research demonstrated that high level features like object proposals have high potential in VPR [13]. In order to extract ROIs, max-pooling on cropped areas in CNN layers' features has been utilized in [36]. Multi-scale, non-rigid, pyramidal fusion of local features to improve VPR has been studied in [21]. In a recently proposed approach [5], a global matching-based, less-intensive place candidates selection is followed by local feature-based, more-intensive final candidate collection with focus on spatial constraints. On challenging Places-365 dataset [43], deep CNNs such as VGG-16 and DenseNet-161 are capable of achieving classification accuracies within 55% and 56%, respectively. It is worth noting that the classification performance is lower than performances achieved by those networks on the ImageNet dataset. In this context it is worth mentioning that images acquired by humanoid robots, and in particular drones or body-worn cameras are even more harder to classify.

3. Algorithm and Experimental Setup

At the beginning we propose an algorithm for blur detection. Next, we present minimum spanning tree for image retrieval and distance-based one-class classification. Then, in next Subsection we describe our dataset for learning image deblurring and place recognition. In the last Subsection we present the whole algorithm for place recognition.

3.1. Algorithm for Blur Detection

Image representation is the major part of a visual place recognition system. At the same time, it is very similar to image representations in systems such as image retrieval, object detection, image classification, and so on. The VPR methods to a large extent draw on the best practice that have been developed in the area of image retrieval. Differences and similarities between VPR and the image retrieval are outlined in a recent review [42]. The main difference is that in VPR the camera position should be taken into account such that images with changes induced by, for instance, different views should be returned as irrelevant. Despite considerable research efforts, robust place recognition in indoor

environments on the basis of a body-worn camera or robot's on-board camera is an unsolved problem. On challenging Places-365 dataset the classification accuracies achieved by deep CNNs are lower than accuracies achieved by those networks on ImageNet dataset. The accuracies on real images acquired by moving/rotating cameras are either too low for applications for the visually impaired or are obtained with a computational cost that prevents real-time applications. The most common approach to VPR relies on learning or embedding image features. A recent approach [5] is an example of a different approach, where a global matching-based, less-intensive place selection of candidates is executed first, and then a local feature-based, more-intensive final selection of candidates selection with focus on spatial constraints is executed afterwards. From recent surveys [6, 42] it follows that vast VPR algorithms do not include scenarios with motion blur, including ones when the robot or camera makes considerable inter-frame rotations. Moreover, most of the recent approaches are based on features extracted by AlexNet (AlexNet365) and VGG16 CNNs, which have 60M and 138M parameters, respectively, and VGG-M architectures that have several times less parameters were not of wider interest.

First, we generated a dataset with images contaminated by motion blur. We used MIT Indoor scene database [30] that comprises 15620 images with 67 indoor categories. The number of examples varies across categories, but there are at least 100 images per category. A Matlab function fspecial was used to approximate the linear motion of a camera with provided lengths and directions. Motivated by recent research findings showing that CNN-based description of places or images using only regions of interest (ROI) leads to enhanced performance compared to whole-image description [37] we based our algorithm on such an approach. In [37] the ROI-based vector representation is proposed to encode several image regions with simple aggregation. An approach proposed in [8] employs a late convolutional layer as a landmark detector and a prior one in order to calculate local descriptors for matching such detected landmarks. For such a regions-based feature encoding a $10k$ bag-of-words (BoW) [32] codebook has been utilized.

The proposed approach to blur detection is based on salient CNN-based regional representations. We used a VGG-M neural network trained on ImageNet dataset. The network consists of five convolutional layers. Just over 3.5 G MAC (multiply-accumulate) operations is needed to classify a single RGB image of size $224 \times 224 \times 3$ from the ImageNet dataset. The image classification is done in almost five times shorter time in comparison to time needed by frequently used VGG-16, which requires more than 15.4 G MAC operations. We employed only convolutional layers, i.e. we discarded the fully connected layers. The fifteenth layer has been leveraged for discovering meaningful regions in the image, on the basis of which the local image features from the lower (thirteenth) layer have been determined. The features extracted from discussed layers were defined as $C_{13}, C_{15} \in R^{13\times13\times512}$, respectively. At the beginning, we perform averaging of activations belonging to clusters determined in advance, and which group non-zero and spatially proximal 8-connected activations. Then, we determine a predefined number of clusters with the highest averaged values and identify the corresponding image ROIs, which are the silent regions. This means that in our approach we perform blur detection not on the whole image but instead we employ only salient CNN-based regional representations of the image. As in [8] we use a higher convolutional layer to guide extraction of local features and to create multiple region descriptors representing each image. For each ROI from the predefined number of ROIs we determined a pooled feature vector as its representation.

At the training stage for each image with and without blur we extracted ten descriptors of size equal to 512. We trained a neural network with one hidden layer to classify images into two categories. The number of neurons in the hidden layer was equal to 20. In order to obtain probabilities of respective label the classifier's output has been calibrated. This way the probability determined by the neural network provides a kind of confidence on the prediction. The calibration has been done using algorithm proposed in [40]. The trained neural network has then been utilized to detect (unknown) blurs. For blur detection on each image we determined 200 descriptors. The response of the trained classifier was averaged over the regions and the image patches were classified as blurred or sharp. For visualization purposes they have also been projected onto the input images, see Fig. 1 that depicts example images. The averaged classifier outputs are equal to 0.056, 0.120, 0.330 and 0.585, respectively.

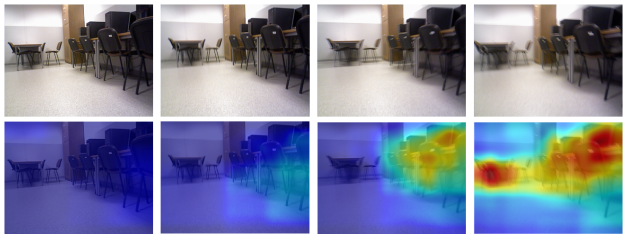

Figure 1: Heat maps of images with increasing blur intensity (top: input images, bottom: corresponding heat maps).

We considered various approaches to quantifying the quality of predictions. We compared the proposed detector with a Support Vector Machine (SVM) with calibrated output as well as Logistic Regression (LR), which returns well calibrated predictions by default as it directly optimizes the logistic regression loss. We experimented with various numbers of descriptor vectors extracted on the test

images. Figure 2 depicts sample images with example number of descriptors that were considered during experiments and evaluations. As we can see, the heat maps vary depending on number of descriptor vectors. Thus, we experimentally determined the number of descriptors leading to best blur detections and then determined the threshold to decide on the basis of averaged predictors of the calibrated classifier if the input image is blurred or sharp one.

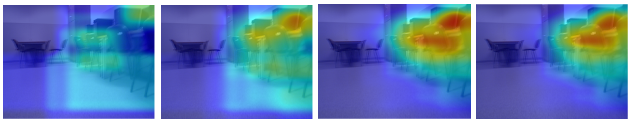

Figure 2: Estimated blur intensity vs. number of descriptor vectors (50, 100, 300 and 400) extracted on blurred image (from 3rd column on Fig. 1).

3.2. Minimum Spanning Tree for Image Retrieval and Distance-based One-class Classification

Many machine learning algorithms have obtained promising results through describing data by graphs such as minimum spanning trees (MSTs). Several clustering algorithms rely on graphs. In our approach, aside of k-NN we investigate also a minimum spanning tree as a class descriptor as well as for image retrieval. A minimum spanning tree is a subset of edges of undirected graph, which connects all vertices jointly, without any cycles and with the minimum total edge weight. Determining the MST is a well-known problem of combinatorial optimization. Conventional minimum spanning tree-based clustering algorithms are known to be capable of detecting clusters with irregular boundaries and overcome many of the problems faced by the classical algorithms [23]. The property that there are no cycles in the tree means that there is only one path among any two nodes. In this work we pre-compute the MST for connecting all images in the training set. The nodes of the tree are connected by the edges and weights express distances between them. They are computed using cosine similarity between global image descriptors. If the sum of the weights of the edges connected by a node is smaller, it means that the redundancy between the corresponding descriptor and relevant image descriptors is lower, and thus the resulting discriminative power of this descriptor is higher, which means that the descriptor is more informative. The classification of test images relies on their distances to the closest edges in the tree and the final decision is taken by a distance-based one-class classifier [14].

3.3. Dataset for Learning Image Deblurring, Blur Recognition and Place Recognition

The past work on VPR focused on developing algorithms operating on images with planar viewpoint changes. Maffra et al. recorded an OldCity dataset [19] that contains walking sequences from the old city of Zurich. More recently, they presented a new dataset as well as an algorithm combining 2D and 3D information for UAV navigation. The discussed datasets were recorded outdoor and do not contain images with unknown blur that can arise during movement of a body-worn camera or even a humanoid robot making rotations about its axis with considerable angular velocity. Moreover, they do not contain information about category or class of visited places. Motivated by lack of datasets with manually labeled blurry images as well as data permitting investigation of approaches based on similarity propagation/diffusion or region manifolds, to name a few promising research directions, we decided to record a new dataset that could fill the existing gap and meet the demand for this type of data. Recently, a real-world blur dataset for learning and benchmarking deblurring algorithms has been proposed [31]. Although the mentioned above dataset meets the real demand for such a real-world data, it does not fill the gap and the need for datasets not only for benchmarking deblurring algorithms, but also for place recognition on images with 6DoF motion. The dataset has been recorded using an RGB camera mounted on the head of a humanoid robot. The dataset includes 10800 RGB images that were acquired by an autonomous robot in nine indoor rooms. Each image has been manually assigned to one of three classes, namely: sharp, blurry and considerably blurry. The training sequence contains 5287 blurry images and 1913 sharp images. Given that images are of size 640×480, image patches can be extracted on such images and then used for training deep neural networks. Two test sequences contain 1366 and 1440 blurred images as well as 434 and 360 sharp images, respectively. We also manually determined twenty four reference images with corresponding relevant and irrelevant images for place recognition. For each place and corresponding query image, several relevant and minimum five irrelevant images were designated, that is, images very similar to the query image as well as relevant images, but differing enough to be considered as relevant. Considerable amount of work of several annotators results in dataset posing several research challenges and opportunities. The dataset can be downloaded from authors' webpages[1] .

3.4. Algorithm

We trained calibrated classifiers to estimate the blur intensity and then use it do detect if the input image is blurry or sharp one. Having on regard that the NetVLAD delivers a powerful pooling mechanism with learnable parameters that can be readily plugged into any other CNN architecture or classifier we trained and then evaluated a set of classifiers for room category recognition. A selected classifier

[1] http://home.agh.edu.pl/~bkw/data/BD-PR/, http://pwozniak.kia.prz.edu.pl

is then used to recognize the room category. We utilized VGG16 and added the NetVLAD layer after the conv_5 layer in order to extract the VLAD features. Given this and other selected features we precalculated the minimum spanning trees and evaluated them for place recognition. At this stage of the research we experimented with various configurations of the algorithm to evaluate the usefulness of blur detection as well as robustness of room classification on the performance of place recognition. We observed that knowledge about motion blur and room category has considerable influence on the final decision because in rooms like corridors the place recognition performance and ability do precisely determine the previously visited place is lower. Finally, a image retrieval engine for determining the most similar image and deciding if it is relevant or irrelevant with the current query image has been developed. In option of the algorithm using the minimum spanning tree, the place recognition can be achieved using high-level information from noise detector, room recognition and information extracted on the basis of the MST. By calculating the distances between descriptor extracted from the current query image and descriptors from the nodes we can quickly determine the relevant sub-tree. Usually, descriptors in the same cluster have similar properties and tend to belong to same class. However, when in the same cluster there are exemplars belonging to different classes then the confidence of final decision is lowered.

The final most relevant image to a given query image is determined dynamically of the basis of two most similar images. This is contrast to relevant algorithms. Lets us assume that for a given query image the most similar images are determined by four different methods, where some of them operate on identical descriptors, whereas some of them operate on different image descriptors. From four images we select two most similar images to the considered query image together with their indexes. Such a pair of indexes defines a series of images that are between two most similar images, selected by two best methods for the considered query image. The final index is simply the average of the indexes of such two most similar images. For such a rounded averaged index to integer value we select the corresponding image that is the final most relevant image. This means that the final decision is taken on the basis of voting.

4. Experimental Results

At the beginning we conducted experiments consisting in motion blur detection as well as deblurring real-world images. We ran our algorithm for blur detection on real images with severe (unknown) blurs and compared it with state-of-the-art algorithms, including [25, 28]. Table 1 presents experimental results that were achieved on two test sequences, where first one is contaminaed by the blur, whereas the second one is more blurry. As we can observe, our algorithm achieves superior results. The images from the second sequence are far more harder to process and to analyze due to higher amount of blur. The results achieved by CNNs specialized for non-uniform blur detection [34] are better in comparison to results achieved on the basis of method [28]. The discussed result has been achieved using the neural network trained in 50 epochs. A recently proposed algorithm [10] achieved accuracy equal to 85.6%.

Afterwards, we determined descriptors representing images and calculated minimum spanning trees. The MSTs were visualized for images from each category as well as for all images from the training set. Figure 3 depicts a sample MST that was obtained for the NetVLAD descriptor on all images from the training subset. As we can observe, in most of subtrees the majority nodes are of the same class. However, in some of them there are nodes from more than one class. Moreover, they are not convex, or of irregular shape, which testify the validity of carrying out the inference on a minimum spanning tree, instead of on a densely connected graph. By definition, the discussed data structure joints the points that are close in the descriptor space, high-lighting intrinsic localities, similarities and local consistencies in the scene. We calculated, visualized and analyzed minimum-spanning trees on all images, on images classified as sharp, and on only blurry images. The discussed analysis of linkage maps was conducted with aim to collect the knowledge about the dataset, to study the usefulness of MST for visual place recognition, and in particular to investigate the influence of blur on performance of scene classification as well as place recognition on images with severe blurs. Particularly, for the selected query images to perform place recognition we analyzed their global descriptors if they are sufficiently rare so that the corresponding locations could be considered distinctive.

Next, we evaluated state-of-the-art global descriptors for indoor scene recognition, where the set of scenes was a list of nine different room types. Table 2 presents experimental results which were achieved on our dataset. We compared the performances achieved by the SVM with the linear kernel as well as k-NN. Table 2 presents only the better result among the results obtained by SVM and k-NN. Classification accuracies achieved on the basis of ReNet50 and SVM are noticeably better in comparison to results achieved on the basis of other deep neural architectures, including GoogleNet trained on Places365 dataset. The classification results achieved by k-NN on NetVLAD features are better in comparison do results mentioned above. The discussed global features have been calculated using VGG-16 + NetVLAD + whitening, trained on Tokyo Time Machine dataset[2] [1]. Although MST did not permit achieving best results, the certainty of the MST-based classifier's decisions is bigger. It is worth emphasizing that categorization

[2] https://www.di.ens.fr/willow/research/netvlad/

Table 1: Blur detection on images with severe (unknown) blurs.

	Seq. with less blur (Seq. #2)				Seq. with more blur (Seq. #3)			
method	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
var. Laplacian	0.8589	0.8114	0.7931	0.8015	0.8767	0.8248	0.7635	0.7880
SVM calib.	0.9078	0.8650	0.7931	0.8992	0.9294	0.8761	0.9226	0.8964
LR	0.9194	0.8984	0.8770	0.8870	0.9228	0.9271	0.8992	0.9065
MB-det-CNN [28]	0.8720	0.8412	0.8231	0.8126	0.8866	0.8398	0.8435	0.8480
BD-PR (ours)	**0.9206**	**0.8869**	**0.9005**	**0.8934**	**0.9428**	**0.9271**	**0.8992**	**0.9065**

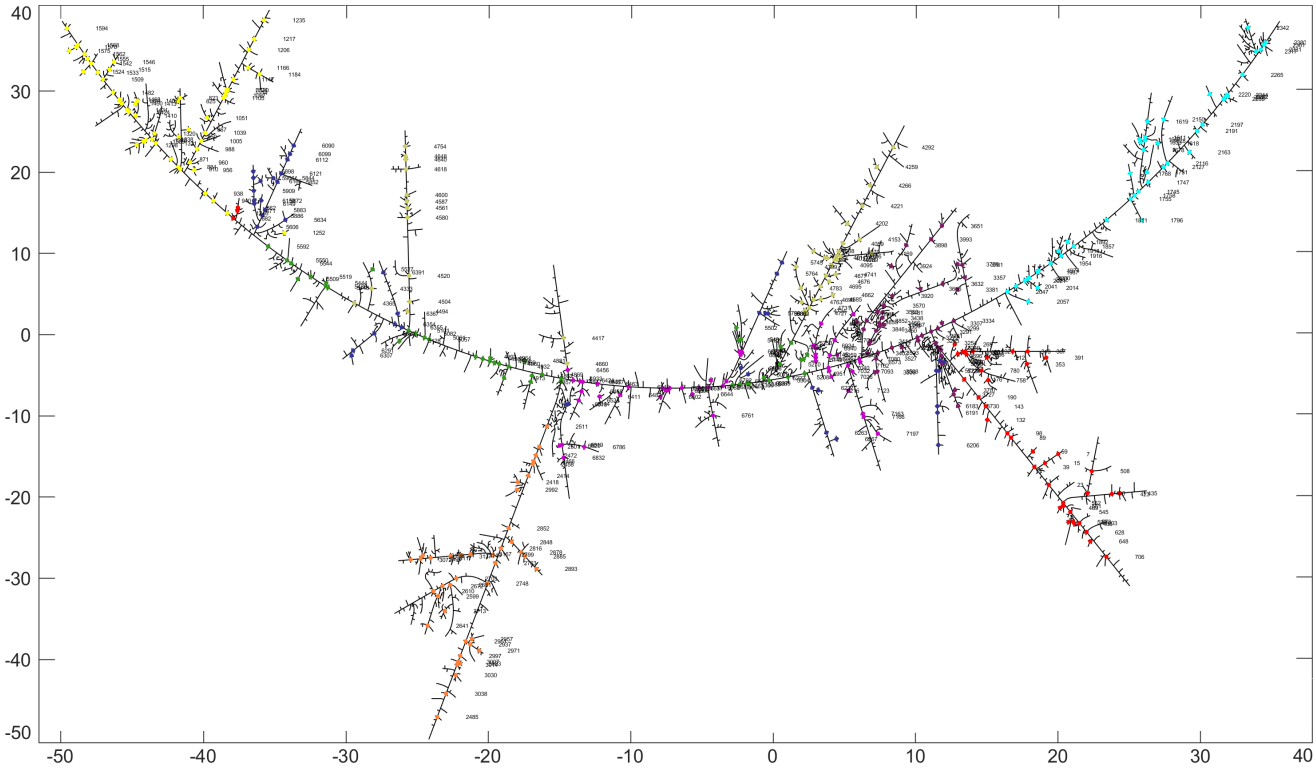

Figure 3: Minimum spanning tree determined on NetLAD descriptor from training subset (plot best viewed in color).

of scenes only on images without blur, i.e. images automatically classified as non-blurry leads to considerable improvement of the results. This means that in such a scenario the algorithm first classifies the acquired image as blurry or non-blurry and then if the image is blurry it acquires next one. As we can observe, costly and time consuming deblurring images with severe (unknown) blurs did not lead to better results. The discussed results were achieved using recently proposed deblurring algorithm [35]. Blur detection and then deblurring the images contaminated by blurs leads only to slightly better results, see results in the last row.

Table 3 presents experimental results achieved on images from another image sequence, on which degree of blur is higher, see also Tab. 1. In general, all scores are much lower than scores presented in Tab. 2. Once again, results achieved on the basis of GoogleNet trained on Place365 dataset are lower in comparison to results achieved on the basis of VGG19 and ResNet50. As we can notice, scene categorization only on images classified as non-blur leads to considerable gain in the classification performance. In contrast to results obtained on images from the image sequence with less noise, the results achieved by algorithm with blur detection and then deblurring the images contaminated by the blur leads to larger gains in the performance in comparison to the case with deblurring all images. This more challenging image sequence of images opens up a number of research opportunities in the field of both blur detection, deblurring images contaminated by severe (unknown) blurs. In particular, comparisons of performances on both sequences can lead to further improvements of algorithms, including algorithms for place recognition.

In last part of experiments we focused on place recog-

Table 2: Performance of room classification on Seq. #2 with less blur (NV - NetVLAD).

	Acc.	Prec.	Recall	F1-sc.
[A] VGG19+SVM	0.9056	0.9072	0.9056	0.9050
[B] GoogleNet Places365+SVM	0.8939	0.8956	0.8939	0.8936
[C] ResNet50+SVM	0.9428	0.9474	0.9428	0.9434
[D] NV+KNN	0.9583	0.9600	0.9583	0.9583
[E] NV+MST	0.9544	0.9567	0.9544	0.9545
[F] NV+SVM+BlurDet.	**0.9652**	**0.9687**	**0.9652**	**0.9662**
[G] NV+SVM+Deblur	0.9528	0.9570	0.9528	0.9532
[H] NV+SVM+BlurDet.+Deblur	0.9550	0.9585	0.9550	0.9556

Table 3: Performance of room classification on Seq. #3 with more blur (NV - NetVLAD).

	Acc.	Prec	Recall	F1-sc.
[A] VGG19+SVM	0.8111	0.8355	0.8111	0.8095
[B] GoogleNet Places365+KNN	0.8056	0.8165	0.8056	0.8053
[C] ResNet50+KNN	0.8444	0.8574	0.8444	0.8458
[D] NV+SVM	0.8678	0.8796	0.8678	0.8690
[E] NV+MST	0.8522	0.8896	0.8522	0.8566
[F] NV+SVM+BlurDet.	**0.9190**	**0.9281**	**0.9135**	**0.9173**
[G] NV+SVM+Deblur	0.8478	0.8655	0.8478	0.8491
[H] NV+SVM+BlurDet.+Deblur	0.8661	0.8832	0.8661	0.8675

nition. As mentioned above, basic idea of current image-based approaches to place recognition is to search a repository of indoor images and return the best match. In the first phase of this part of the research, we analyzed the performance of place recognition on images from the less blurred sequence using the NetVLAD features. The precision is the fraction/percentage of retrieved images that are relevant. The recall is the fraction/percentage of relevant images that were retrieved. Figure 4 depicts sample precision-recall plot for one of the rooms (F104). Minimum two query places for each room with corresponding relevant and irrelevant images were manually selected for evaluation of algorithms for visual place recognition.

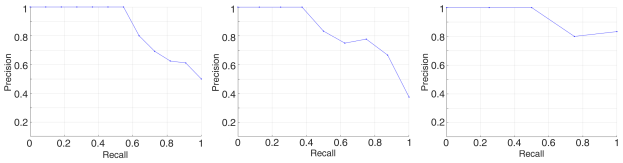

Figure 4: Precision-recall plots for three query images from room F104.

Figure 5 depicts precision-recall plots for all rooms using images from sequence with less blur (Seq. #2). The red horizontal lines depict the mean Average Precision (mAP) values for the considered scenes. It can be seen that the performance of the algorithm is lower for the corridors, see plots in the first row, while the performance of discussed algorithm is acceptable for all remaining rooms.

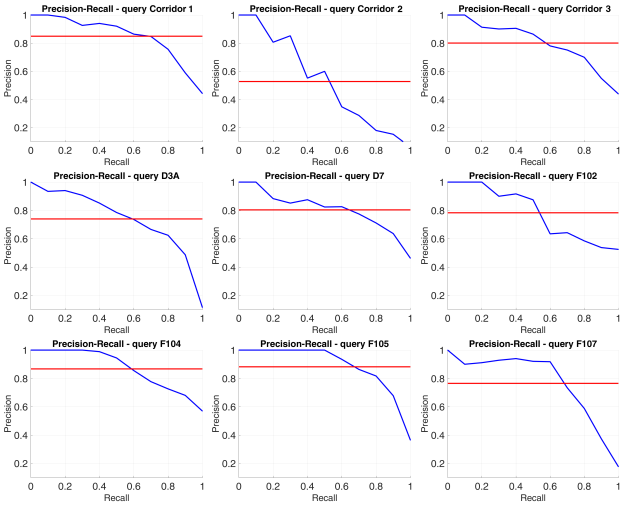

Figure 5: Precision-recall plots for all nine scenes.

First row of Figure 6 depicts query image and then relevant images, which are sorted from most similar to less similar. Second row contains example irrelevant images. The discussed images, except query one, were manually selected taking into account the perceptual similarity/dissimilarity with the query image. Third row shows some correctly matched query and reference images.

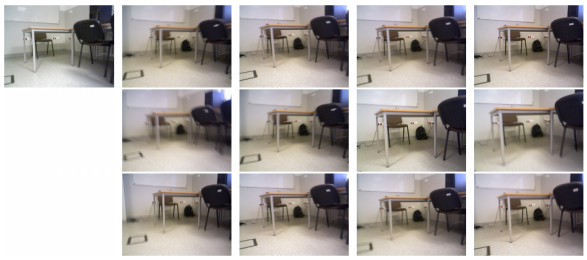

Figure 6: Query image and relevant images (upper row), irrelevant images (second row), images retrieved using NetVLAD features.

In the last phase of this part of the research we compared performances of algorithms in place recognition. We evaluated recognition performance achieved by k-NN (k set to 3) on netVLAD features with VGG-M as backbone, MST on netVLAD features with VGG-M as backbone, onlyLookOnce (oLN) [8] with VGG-M as backbone, and correlations between Gram matrixes extracted on conv5_3 layers of VGG-M network. Each of the mentioned above image retrieval methods was evaluated on all images or images on which no blur has been detected. Table 4 illustrates the mean Average Precision (mAP) for each of the considered scenes as well as the mAP scores for all scenes and images from Seq. #2. In the last two columns we present results that were achieved on the basis of the proposed voting.

Table 4: mAP achieved by k-NN (VGG-M, NetVLAD), MST (VGG-M, NetVLAD), onlyLookOnce (VGG-M), correlation between Gram matrixes (VGG-M) and our voting-based approach, with no blur detection and blur detection (bd.).

	room	k-NN	k-NN+bd.	MST	MST+bd.	oLN	oLN+bd.	GM	GM+bd.	voting	voting+bd.
1	Corr. 1	0.8643	1.0	0.8625	0.9377	0.9552	1.0	0.8011	0.9931	0.9425	1.0
2	Corr. 2	0.5804	1.0	0.5789	1.0	0.8573	1.0	0.9084	1.0	0.6864	1.0
3	Corr. 3	0.8007	0.9750	0.7945	0.9750	0.8270	0.7667	0.7045	0.7917	0.8629	0.9750
4	D3A	0.7831	0.6549	0.7836	0.6549	0.8482	0.6913	0.8243	0.7931	0.8336	0.6987
5	D7	0.8038	0.8193	0.8016	0.8193	0.8083	0.8788	0.8985	0.9859	0.8698	0.8762
6	F102	0.7833	1.0	0.7146	1.0	0.9406	1.0	0.8559	1.0	0.8329	1.0
7	F104	0.8704	0.8547	0.8504	0.8537	0.8483	0.8543	0.6687	0.9103	0.8652	0.9004
8	F105	0.8819	1.0	0.8813	1.0	0.9123	0.9167	0.7737	0.7421	0.9172	1.0
9	F107	0.7238	0.8772	0.7238	0.8772	0.7663	0.8600	0.4711	0.5184	0.7620	0.8551
	mAP	0.7880	0.9090	0.7768	0.9020	0.8626	0.8864	0.7674	0.8594	0.8414	**0.9228**

In general, blur detection permits to achieve considerable better mAP scores, c.f. results in the last row in Tab. 4. Comparing results achieved by k-NN operating on netVLAD features with VGG-M as backbone, we can observe that thanks to blur detection, considerable gains in the mAP scores can be obtained.

As we can notice, the discussed scores vary considerably depending on the room category. The proposed correlation between Gram matrixes, which are extracted on conv5_3 layers of VGG-M network, achieved better mAP scores in comparison to scores of k-NN in room #2 (corridor), and rooms #4, #5 and #6. As in the case of the k-NN, blur detection allows to achieve far fetter results. It is worth noting that onlyLookOnce is quite resistant to motion blur as gains in mAP scores when using blur detection to skip images with blur are relatively small. It transpired that knowledge about room category is essential as confidence of place recognition in long and narrow corridors with similar and repetitive scene content is much smaller. MST-based analysis of neighbors with respect to coherence of global descriptors, together with information about image noise as well as room category can improve the performance of place recognition. As we can observe in Tab. 4 the use of such information in the voting engine permits to achieve the best performance of place recognition in terms of mAP.

5. Conclusions

In this work we introduce a challenging dataset for blur detection and visual place recognition. The RGB images were acquired by a humanoid robot in nine rooms. We present a novel real-time algorithm for blur detection on images with severe (unknown) motion blur and demonstrate experimentally that it outperforms recent algorithms. Owing to using a computationally cheap neural backbone for regional feature extraction it can be executed in real-time on embedded devices, including recently introduced by Intel device for visually impaired. Local features that are calculated on salient CNN-based regional representations are fed to calibrated classifiers in order to estimate the blur intensity. We evaluate potential of several global descriptors for scene classification and demonstrate experimentally that blur detection permits to achieve superior accuracies in room recognition. We propose an algorithm for place recognition that on the basis of two most relevant images, which are selected by two best methods established dynamically, determines the final most relevant image. We demonstrate experimentally that such an approach to visual place recognition permits achieving superior mean average precision.

Acknowledgement

This work was supported by Polish National Science Center (NCN) under a research grant 2017/27/B/ST6/01743.

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. *PAMI*, 40(6):1437–1451, 2018. 2, 5

[2] R. Arandjelovic and A. Zisserman. All About VLAD. In *CVPR*, pages 1578–1585, 2013. 2

[3] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit. Instant outdoor localization and SLAM initialization from 2.5d maps. *IEEE TVCG*, 21(11):1309–1318, 2015. 2

[4] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. Leonard. Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age. *IEEE Tr. on Rob.*, 32:1309–1332, 2016. 1

[5] L. Camara and L. Peuil. Visual place recognition by spatial matching of high-level CNN features. *Robotics and Autonomous Systems*, 133:103625, 2020. 2, 3

[6] S. Cebollada, L. Pay, M. Flores, A. Peidr, and O. Reinoso. A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. *Expert Syst. with Appl.*, pages 114–195, 2020. 1, 3

[7] Z. Chen, O. Lam, A. Jacobson, and M. Milford. Convolutional neural network-based place recognition. In *Proc. Austral. Conf. on Rob. and Aut.*, pages 1–8, 2014. 2

[8] Z. Chen, F. Maffra, I. Sa, and M. Chli. Only look once, mining distinctive landmarks from ConvNet for visual place recognition. In *IROS*, pages 9–16, 2017. 3, 7

[9] M. Cummins and P. Newman. Appearance-only SLAM at large scale with FAB-MAP 2.0. *Int. J. Rob. Res.*, 30(9):1100–1123, 2011. 2

[10] X. Cun and C.-M. Pun. Defocus blur detection via depth distillation. In *ECCV*, pages 747–763, 2020. 5

[11] Y. Fang, K. Wang, R. Cheng, K. Yang, and J. Bai. Visual place recognition based on multilevel descriptors for the visually impaired people. In *Target and Background Signatures V*, volume 11158, pages 21 – 32. SPIE, 2019. 2

[12] S. Garg and M. Milford. Straightening sequence-search for appearance-invariant place recognition using robust motion estimation. In *Proc. Austral. Conf. on Rob. and Aut.*, pages 203–212, 2017. 1

[13] Y. Hou, H. Zhang, and S. Zhou. Evaluation of object proposals and ConvNet features for landmark-based visual place recognition. *J. Intell. & Rob. Syst.*, 92(3-4):505–520, 2017. 2

[14] P. Juszczak, D. Tax, E. Pekalska, and R. Duin. Minimum spanning tree based one-class classifier. *Neurocomputing*, 72(7):1859–1869, 2009. 4

[15] T. Levitt and D. Lawton. Qualitative navigation for mobile robots. *Artificial Intell.*, 44(3):305 – 360, 1990. 1

[16] A. Lopez-Cifuentes, M. Escudero-Vinolo, J. Bescoos, and A. Garcia-Martin. Semantic-aware scene recognition. *Pattern Rec.*, 102:107256, 2020. 2

[17] S. Lowry, N. Suenderhauf, P. Newman, J. Leonard, D. Cox, P. Corke, and M. Milford. Visual place recognition: A survey. *IEEE Trans. on Robotics*, 32:1–19, 2016. 1, 2

[18] F. Maffra, Z. Chen, and M. Chli. Viewpoint-tolerant place recognition combining 2D and 3D information for UAV navigation. In *ICRA*, pages 2542–2549, 2018. 1

[19] F. Maffra, L. Teixeira, Z. Chen, and M. Chli. Loop-closure detection in urban scenes for autonomous robot navigation. In *Int. Conf. on 3D Vision (3DV)*, pages 356–364, 2017. 4

[20] J. K. Mahendran. Case study: A vision system for the visually impaired. Technical report, Intel, March 2021. 2

[21] J. Mao, X. Hu, X. He, L. Zhang, L. Wu, and M. Milford. Learning to fuse multiscale features for visual place recognition. *IEEE Access*, 7:5723–5735, 2019. 2

[22] M. J. Milford and G. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *ICRA*, pages 1643–1649, 2012. 2

[23] G. Mishra and S. Mohanty. A fast hybrid clustering technique based on local nearest neighbor using minimum spanning tree. *Expert Syst. with Appl.*, 132:28–43, 2019. 4

[24] L. Murphy and G. Sibley. Incremental unsupervised topological place discovery. In *ICRA*, pages 1312–1318, 2014. 2

[25] N. D. Narvekar and L. J. Karam. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Tr. Image Proc.*, 20(9):2678–2683, 2011. 5

[26] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int J. of Comp. Vis.*, 42(3):145–175, 2001. 2

[27] E. Ovalle-Magallanes, N. Aldana-Murillo, J. Avina-Cervantes, J. Ruiz-Pinales, J. Cepeda-Negrete, and S. Ledesma. Transfer learning for humanoid robot appearance-based localization in a visual map. *IEEE Access*, 9:6868–6877, 2021. 1

[28] J. L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. Diatom autofocusing in brightfield microscopy: A comparative study. In *Proc. 15th Int. Conf. on Pattern Rec.*, volume 3, pages 314–317, 2000. 5, 6

[29] A. Pretto, E. Menegatti, M. Bennewitz, W. Burgard, and E. Pagello. A visual odometry framework robust to motion blur. In *ICRA*, pages 2250–2257, 2009. 1

[30] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, pages 413–420, 2009. 3

[31] J. Rim, H. Lee, J. Won, and S. Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201, 2020. 4

[32] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *IEEE Int. Conf. on Comp. Vis.*, pages 1470–1477, 2003. 3

[33] N. Suenderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. On the performance of ConvNet features for place recognition. In *IROS*, pages 4297–4304, 2015. 2

[34] J. Sun, W. Cao, Z. Xu, and J. Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *CVPR*, pages 769–777, 2015. 5

[35] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 6

[36] G. Tolias, Y. Avrithis, and H. Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *IEEE Int. Conf. on Comp. Vis.*, pages 1401–1408, 2013. 2

[37] G. Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of CNN activations. In *Int. Conf. on Learning Repr.*, 2016. 3

[38] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, pages 1808–1817, 2015. 1

[39] A. Yandex and V. Lempitsky. Aggregating local deep features for image retrieval. In *IEEE Int. Conf. on Comp. Vis.*, pages 1269–1277, 2015. 2

[40] B. Zadrozny and Ch. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, pages 694–699. ACM, 2002. 3

[41] M. Zaffar, A. Khaliq, S. Ehsan, M. Milford, and K. McDonald-Maier. Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions. *CoRR*, abs/1207.0016, 2019. 2

[42] Xiwu Zhang, Lei Wang, and Yan Su. Visual place recognition: A survey from deep learning perspective. *Pattern Rec.*, page 107760, 2020. 2, 3

[43] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 40(6):1452–1464, 2018. 2