# Place Inference via Graph-based Decisions on Deep Embeddings and Blur Detections

Piotr Wozniak[2] and Bogdan Kwolek[1]✉

[1] AGH University of Science and Technology, 30 Mickiewicza, 30-059 Kraków, Poland
http://home.agh.edu.pl/~bkw/contact.html
[2] Rzeszów University of Technology
Al. Powstańców Warszawy 12, 35-959 Rzeszów, Poland

**Abstract.** Current approaches to visual place recognition for loop closure do not provide information about confidence of decisions. In this work we present an algorithm for place recognition on the basis of graph-based decisions on deep embeddings and blur detections. The graph constructed in advance permits together with information about the room category an inference on usefulness of place recognition, and in particular, it enables the evaluation the confidence of final decision. We demonstrate experimentally that thanks to proposed blur detection the accuracy of scene recognition is much higher. We evaluate performance of place recognition on the basis of manually selected places for recognition with corresponding sets of relevant and irrelevant images. The algorithm has been evaluated on large dataset for visual place recognition that contains both images with severe (unknown) blurs and sharp images. Images with 6-DOF viewpoint variations were recorded using a humanoid robot.

**Keywords:** Visual place recognition, CNNs, images with unknown blur

## 1 Introduction

Simultaneous localization and mapping (SLAM) is the computational problem aiming at constructing and updating a map of an unknown environment while simultaneously keeping path of an agent's location within it [1]. Although SLAM is used in many practical applications, several challenges prevent its wider adoption. Since SLAM is based on sequential movement and measurements that are contaminated by some margin of error, the error accumulates over time, causing substantial deviation from actual agent's locations. This can in turn lead to map distortion or even collapse and thus making subsequent searches difficult. Loop closure is a task consisting in recognition of previously-visited location and updating the constructed map accordingly. Therefore, detecting loop closure (or previously visited places) in order to correct the accumulated error during the exploration is very important task [2]. This permits the SLAM system to relocalize the sensor after a tracking failure, which might happen in unfavorable circumstances, like severe occlusion or abrupt movements.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

The aim of visual place recognition (VPR) is to retrieve correct place matches under viewpoint and illumination variations, while requiring as less as possible computational power and memory storage [3]. Over the past years several methods for visual place recognition have been developed [3,2]. Although most of visual place recognition methods were developed for SLAM, VPR algorithms also found applications in monitoring of electricity pylons using aerial imagery [4], brain-inspired navigation [5], and image-search based on visual content [6]. VPR is very challenging problem because images of the same place but taken at different times may differ notably from each other. The differences can be caused by factors such as varying illumination, shadows as well as changes resulting from different passing the same route.

In robotics, most of evaluations of VPR systems were performed using data acquired by ground-based mobile platforms or robots. The degree of viewpoint variation that takes place during scene perception by a humanoid robot is far more complex than viewpoint variations experienced by mobile robots [7]. When a humanoid robot is walking, turning, or squatting, its head mounted camera moves in a jerky and sometimes unpredictable way [8]. Motion blur, one of the biggest problems for feature-based SLAM systems, causes inaccuracies and location losses during map construction. Most of datasets for visual place recognition provide lateral or 3D variations of viewpoint. The 24/7 Query dataset [9] contains outdoor images with 6-DOF viewpoint variation. Recently, the Shopping street dataset targeted for aerial place recognition with 6-DOF viewpoint change has been introduced in [10]. Most of VPR benchmark data are time-based, as frames are acquired and stored at a fixed FPS (frames per second) rate of a video camera. Typically, they are recorded under assumption of non-zero speed of the robot. In [11] a frame is picked every few meters to represent a new place. A disadvantage of both time- and distance-based approaches are huge requirements for data storage. Moreover, they lead to visually similar frames at different places and thus to inaccuracies and impracticality for long-term robot missions.

VPR is typically cast as image retrieval problem. Several handcrafted local and global feature descriptors were proposed for place recognition [3]. CNNs for visual place recognition were proposed in [12]. Since publication of this seminal work, more and more data-driven image description approaches have emerged. Performance of these algorithms has been studied in [13]. In [14], a VLAD [15] layer for CNN architecture that could be trained in end-to-end fashion, specifically for place recognition task has been proposed. The experimental results achieved by NetVLAD on very challenging datasets significantly outperformed results achieved by pre-trained CNNs. Very high potential of VLAD has recently been confirmed in [16], where a comprehensive comparison of 10 VPR systems identified the NetVLAD as the best overall performing method.

Motivated by lack of a dataset with variations arising during typical movement of humanoid and walking robots, particularly containing images with severe (unknown) blurs we recorded a dataset using camera mounted on head of humanoid robot. To cope with place recognition on the basis of images with unknown blur we propose an effective algorithm for blur detection. We demonstrate

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

experimentally that the proposed algorithm considerably outperforms state-of-the-art algorithms on images with severe and unknown motion blur. We demonstrate also that owing to use of the proposed algorithm, considerable gains of performance in scene categorization can be achieved. We employ minimum spanning tree (MST) for place recognition purposes and show its usefulness. Thanks to information extracted on the basis of MST like proportion of images belonging to given class with respect to number of images from remaining classes in a given tree branch the system can infer about confidence of place recognition.

## 2 Relevant work

Scene recognition is very challenging problem [17,18] and variety of approaches have been proposed during the last years. The most frequently used hand-crafted global descriptor is GIST [19]. With the rise of deep learning, learned features become increasingly widely used in localization algorithms. This resulted in a paradigm shift in VPR research consisting in focusing on neural network activations-based descriptors. Considerable potential of features extracted from CNN layers and used as global descriptors has been demonstrated in [20]. Scale Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF) are two of the most commonly used local descriptors [21]. These local techniques extract invariant keypoints from an image and provide descriptions of these keypoints by an underlying low-level gradient-based descriptors. They have been applied in several algorithms for visual place recognition [3]. However, as observed in [9], SIFT can cope with large changes in appearance and illumination, but only when there is no large view point change [22]. On the other hand, geometric features like vertical lines can be very useful to represent buildings [23] or objects like doors in outdoor/indoor environments.

Pretrained CNN-based approaches to VPR can be roughly divided into two main categories in which: (i) responses from convolutional layers are extracted on the basis of the entire image [12], (ii) salient regions are identified through distinguishing patterns on the basis of convolutional layers responses to entire image [24]. High level features like object proposals have demonstrated remarkable potential in VPR [25]. Philbin et al. [26] learn a non-linear transformation model for descriptors that leads to greatly better matching performance. Tolias et al. [27] use max-pooling on cropped areas in CNN layers' features in order to extract ROIs. Mao et al. [28] propose multi-scale, non-rigid, pyramidal fusion of local features to improve VPR. In [29] a global matching-based, less-intensive place candidates selection is followed by local feature-based, more-intensive final candidate selection with focus on spatial constraints. Deep neural networks such as GoogLeNet, ResNet-152, VGG-16 and DenseNet-161 achieved classification accuracies of 53.6%, 54.7%, 55.2% and 56.1%, respectively on challenging Places-365 dataset [17]. The classification accuracies are lower in comparison to accuracies achieved by those networks on ImageNet dataset. The images acquired by mobile robots, and in particular humanoid robots or drones are even harder to classify. In [30], the transfer learning technique to retrain the VGG-F

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

network in order to categorize places among 16 rooms on images acquired by a humanoid robot has been discussed.

## 3 Algorithm and Experimental Setup

At the beginning of this Section we propose an algorithm for blur detection. Afterwards, we present minimum spanning tree for place recognition. Then, in the next Subsection we describe our dataset. In the last Subsection we present the whole algorithm for place recognition.

### 3.1 Blur Detection

The basic idea of current approaches in robotics to visual place recognition is to search a database of indoor images and return the best match. Considerable attention is devoted to algorithms trained in end-to-end manner. Despite considerable research efforts, robust place recognition in indoor environments on the basis of on-board robot camera is an unsolved problem. The classification accuracies achieved by deep neural networks on challenging Places-365 dataset are lower in comparison to accuracies achieved by those networks on ImageNet dataset. The accuracies on real images acquired during robot motion are either too low for the purposes of loop-closure or are obtained with a high computational cost that prevents real-time applications. A dominating approach consists in learning or embedding features. One of the exceptions is a recent approach [29] in which a global matching-based, less-intensive place candidates selection is realized in advance, and then a local feature-based, more-intensive final candidate selection with focus on spatial constraints is executed. It is also worth noting that most of the approaches to visual place recognition do not consider scenarios with significant motion blur or, as a last resort, neglect motion blur, especially when the robot or camera rotates.

At the beginning we generated a dataset with images contaminated by motion blur. We employed MIT Indoor scene database [31] that consists of 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 has been used to approximate the linear motion of a camera with provided lengths $(5, \ldots, 10)$ and directions $(0, \ldots, \pi/2)$. 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 [32] we based our algorithm on such an approach. In [32] the ROI-based vector representation is proposed to encode several image regions with simple aggregation. An approach proposed in [24] 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) [33] codebook has been utilized. The proposed approach to blur detection is based on salient CNN-based regional representations. The layers conv5_3 and conv5_2 of VGG-16, pre-trained on ImageNet dataset

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

were used to extract the features representing 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 [24] we utilize a higher convolutional layer to guide extraction of local features and to create multiple region descriptors representing each image. At the training stage for each image with and without blur we extracted ten descriptors of size equal to 512, representing image regions with highest average activations. We trained a neural network with one hidden layer to classify the mentioned above image descriptors into two categories. The number of neurons in the hidden layer was equal to 20. The trained neural network has then been used to detect the noise. In testing stage for each image we extracted 200 descriptors as a representation of image regions with highest average activations. The responses of the neural network for such descriptors were averaged. The average values were then used to label the images as blurred or sharp. For visualization purposes the outputs of the classifiers were also projected onto the input images, see Fig. 1 that depicts sample images. For the discussed images the averaged outputs are equal to 0.1476, 0.5333 and 0.8532, respectively.

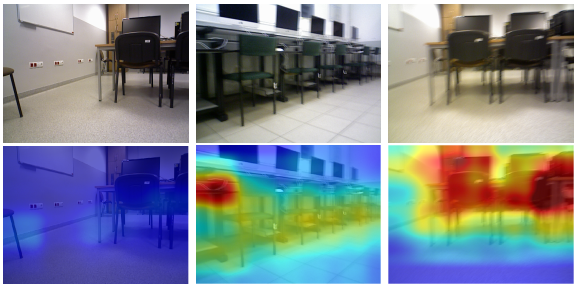

Fig. 1: Heat maps of images with increasing blur intensity.

We experimented with various numbers of descriptor vectors extracted on the test images. Figure 2 depicts sample images with some considered number of descriptors. As we can observe, the depicted heat maps change 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 if image is blurred or sharp one. This problem is an example of multi-objective optimization and in a future work the trade-off between number of descriptors and noise level will be determined automatically.

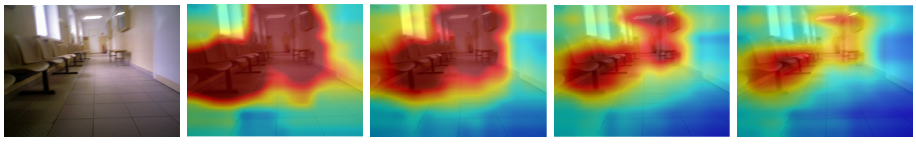

Fig. 2: Blurry input image (left) and heat maps for various number of descriptor vectors (50, 100, 200 and 300) extracted on the blurred image.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

### 3.2 Minimum Spanning Tree-Based Place Recognition

By constructing a minimum-spanning tree the original dense graph is simplified into a minimum weight subgraph, which greatly reduces the number of edges and provides subgraphs of vertices of different degrees. Conventional minimum spanning tree-based clustering algorithms employ information about edges contained in the tree to partition a data set. A minimum spanning tree is a subset of edges of undirected graph that connects all vertices together, without any cycles and with the minimum total edge weight [34]. The property that there are no cycles means that there is only one path among any two nodes in the tree. In this work we compute a MST that connects all images of the training set. Nodes are connected by edges while weights express similarities between them. The edges were determined upon cosine similarity between global descriptors of images. In the proposed approach the MST has been utilized to support the place recognition. For each landmark place a number of relevant images has been determined. The MST has been built upon a selected global descriptor of the images. Given a MST created in advance on the training dataset, for each new image the algorithm seeks for the MST edge that is closest to this new image. The query images were classified as relevant or irrelevant on the basis of their similarities with the closest edges of the tree. Additional information about the room as well as blur of the images has been considered to enhance the place recognition. Moreover, a confidence of the place recognition has also been estimated.

### 3.3 The Dataset

The dataset has been recorded using a RGB camera mounted on the head of a humanoid robot. The dataset contains 9000 images, which were acquired in nine indoor rooms. Each image has been manually classified as sharp or blurred or considerably blurred. The training sequence contains 5287 blurred images and 1913 sharp images. A test sequence contains 1366 blurred images as well as 434 sharp images. For place recognition we also manually determined twenty two reference images with corresponding relevant and irrelevant images.

### 3.4 Algorithm

We trained the neural network to estimate the blur intensity and then used its outputs to detect if the input image is blurry or sharp one. Having on regard that the NetVLAD offers a powerful pooling mechanism with learnable parameters that can be easily plugged into any other CNN architecture or classifier we trained and then evaluated a set of classifiers for room recognition. A selected classifier is then used to recognize the room. 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.

Given all $N$ training images and global descriptors, a pairwise similarity matrix of size $N \times N$ is determined for each descriptor. Afterwards, a MST is

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

built on NetVLAD descriptor. The edges are determined upon cosine similarity between global descriptors of images. Then, blur information as well as room class are included in nodes of the MST built on the NetVLAD. Subsequently, the stored MST tree is processed using query images. Given a query image, only nodes of degree higher than two are assessed with respect to similarity with the query descriptor. Only 0.3 of the most similar nodes with the query descriptor are retained for further analysis. Afterwards, on the basis of the NetVLAD the most similar forty descriptors to the query descriptor together with corresponding node information are selected. Only nodes labeled as sharp as well as with the same class as the query image are included in the subset mentioned above. Such descriptors (images) are then sorted with respect to similarity with the query descriptor (image). Three sorted lists of images are determined for the NetVLAD descriptor and two additional global descriptors. Finally, the order of the images is updated upon the similarities of three global descriptors with the query image. As a result, the two descriptors (for instance Resnet-50 and GoogleNet), which individually get worse results than NetVLAD, in tandem may provide more relevant images to the query images and thus improve the average precision (AP) score of place recognition for a given query image.

Let us assume that we have a sorted list of similarities between the NetVLAD descriptors for the query image and the most relevant images. Let us also assume that we have also an ordered list of similarities between the ResNet50 descriptors and the most relevant images as well as ordered list of similarities between the GoogleNet descriptors and the most relevant images. For the image corresponding to the most similar NetVLAD descriptor with the query descriptor we determine the positions (indexes) in the ordered lists of ResNet50 and GoogleNet descriptors, which were determined for this considered image. We repeat this operation for the remaining descriptors and store indexes in subsequent rows of three column table. After computing the averages for all rows we obtain values which are used to reorder the relevant images with the query image.

We experimented with various configurations of the algorithm to evaluate the usefulness of blur detection as well as influence of classification scores 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 for loop closure is lower. Finally, a classifier built on the MST has been utilized in image retrieval for the most similar image. This means that final decision is taken using high-level information from noise detector, room recognition and information extracted on the basis of the MST.

By calculating the similarity measures between descriptor extracted from the current image and descriptors from the edges we can quickly determine the relevant sub-tree. Usually, descriptors in the same cluster have similar properties and tend to be in the same class. However, when in the same cluster there are exemplars belonging to different classes then the confidence of final decision is lowered. In our approach the confidence of place recognition is determined using the most relevant image found in the place recognition. Using the global descrip-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

tor of this image we searched for fifty most similar images. Such a pool of the most similar images has been determined on the basis of the MST edges holding cosine similarities between NetVLAD descriptors. When the decision confidence is below a threshold it is marked as not valuable for the loop-closure. In the basic approach the confidence has been determined as the ratio of sharp images to total number of images in the pool. We investigated also approaches combining blur information with class information. The MST have been calculated using dd tools [35]. Aside of the NetVLAD we employed the descriptors extracted from Resnet-50 and GoogleNet backbones.

## 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 [36,37]. Table 1 presents experimental results that were achieved on test sequence Seq. #2 from our dataset. As we can observe, the best results were achieved by our algorithm. Taking into consideration that the decision whether the image is sharp is done on the basis of averaging the classifier output we evaluated also SVM with the calibrated output as well as the logistic regression (LR), which generates the calibrated output by default. It is also worth mentioning that the results achieved by CNNs specialized for non-uniform blur detection [38] are better in comparison to results achieved on the basis of method [37]. The discussed result has been achieved using neural network trained in 50 epochs. It has been trained on about 250 000 image descriptors randomly selected from the whole pool of training descriptors, whereas SVM and LR classifiers were trained on 50 and 150 thousand of descriptors, respectively. A recently proposed algorithm [39] achieved accuracy equal to 85.6%.

Table 1: Blur detection on images from Seq. #2 with severe (unknown) blur.

| method | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| var. Laplacian | 0.8589 | 0.8114 | 0.7931 | 0.8015 |
| SVM calibrated | 0.9078 | 0.8650 | 0.8992 | 0.8798 |
| Logistic regression | 0.9194 | 0.8984 | 0.8770 | 0.8870 |
| MB-det-CNN [37] | 0.8720 | 0.8412 | 0.8231 | 0.8126 |
| Our method | **0.9206** | **0.8869** | **0.9005** | **0.8934** |

Afterwards, we determined descriptors representing images and calculated minimum spanning trees. The MSTs were visualized for images from each category as well as all images from the training set. Figure 3 depicts a sample MST that was obtained on the NetVLAD descriptor on all images from the training subset. We calculated, visualized and analyzed minimum-spanning trees on all images, images classified as sharp, and only blurry images. The discussed analysis of linkage maps was conducted with aim to collect the knowledge about

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

dataset, and in particular to investigate influence of the blur on the performance of scene classification as well as place recognition on images with severe blurs.

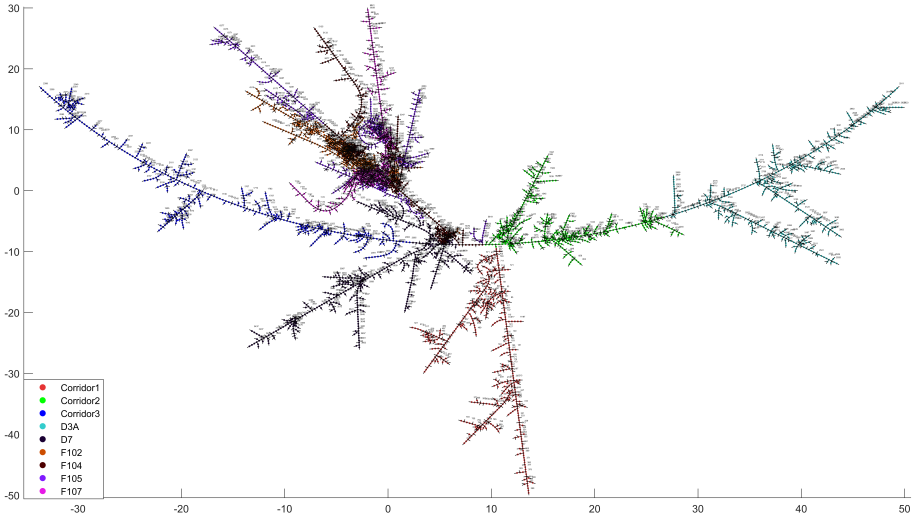


Fig. 3: Minimum spanning tree determined on NetLAD descriptor from training subset (plot best viewed in color).

Next, we evaluated state-of-the-art global descriptors in 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 sequence #2 from our dataset. We compared the performances achieved by the SVM with the linear kernel as well as k-NN. Table 2 presents only better result for each considered case. As we can observe, the categorization performance achieved on the basis of HOG and LBP descriptors is worse in comparison to remaining results. Classification performances achieved in transfer-learning based approach [30] are far better, see results C-E. Moreover, accuracies achieved upon the ReNet50 and SVM are noticeably better in comparison to results achieved on the basis of other deep neural architectures, including GoogleNet trained on Places-365 dataset. The classification results achieved by the k-NN on NetVLAD features are better in comparison to results mentioned above. The features were calculated using VGG-16, NetVLAD with whitening, trained on Tokyo Time Machine dataset [14] (downloaded from `https://www.di.ens.fr/willow/research/netvlad/`). The recognition of rooms 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 robot first classifies the acquired image as blurry or non-blurry and then in case 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 [40]. Blur detection and then deblurring the images contaminated by blurs leads only to slightly better results, see results in the last row.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

Table 2: Performance of room categorization on Seq. #2 from our dataset.

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| [A]HOG+SVM | 0.6872 | 0.7063 | 0.6872 | 0.6921 |
| [B]LBP+SVM | 0.7639 | 0.7867 | 0.7639 | 0.7655 |
| [C]VGG19+SVM | 0.9056 | 0.9072 | 0.9056 | 0.9050 |
| [D]GoogleNet Places-365+SVM | 0.8939 | 0.8956 | 0.8939 | 0.8936 |
| [E]ResNet50+SVM | 0.9428 | 0.9474 | 0.9428 | 0.9434 |
| [F]NetVLAD+KNN | 0.9583 | 0.9600 | 0.9583 | 0.9583 |
| [G]NetVLAD+MST | 0.9544 | 0.9567 | 0.9544 | 0.9545 |
| [H]NetVLAD+SVM+BlurDet. | **0.9652** | **0.9687** | **0.9652** | **0.9662** |
| [I]NetVLAD+SVM+Deblur | 0.9528 | 0.9570 | 0.9528 | 0.9532 |
| [J]NetVLAD+SVM+BlurDet.+Deblur | 0.9550 | 0.9585 | 0.9550 | 0.9556 |

In last part of experiments we focused on place recognition. 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 Seq. #2 using the NetVLAD, GoogleNet and ResNet50 features. The NetVLAD features have been extracted using VGG-M network trained on TokyoTM dataset. The size of the feature vector extracted upon conv5_3 layer is $1 \times 4096$. We utilized GoogleNet trained on Places-365 dataset and ResNet50 trained on the ImageNet. The size of the GoogleNet-based feature vector is $1 \times 1024$ and it was extracted from pool5-7x7_s1 layer. The ResNet50-based feature is of size $1 \times 2048$ and it was extracted from GlobalAveragePooling2DLayer, avg_pool layers. Table 3 presents mean average precision (mAP) scores as well as their average values, which were achieved in recognition of 22 places in nine rooms. The last two columns of the table contain the results that were achieved using the MST and a combined descriptor. For each descriptor we determined the pairwise similarity matrix. The similarity matrixes have then been normalized to 0-1 range. Afterwards an average similarity matrix for all descriptors has been calculated. Finally, we determined the MST of a complete undirected graph with weights given by the averaged similarity matrix. As we can observe, such an algorithm achieved the best mAP scores. Thanks to considering information about blur far better mAP scores can be obtained in place recognition.

Figure 4 depicts precision-recall plots for selected rooms. The precision is the fraction/percentage of retrieved images that are relevant. The recall is the fraction/percentage of relevant images that were retrieved. For the analyzed rooms: D3A, D7, F104 and F107 the number of landmark points was equal to three. For the remaining rooms the precision-recall curves were perfect.

First row of Figure 5 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 perceptual similarity/dissimilarity with the query image. Third

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

Table 3: Performance of place recognition (bd. - blur detection).

| | k-NN VGG-M NetVLAD | | k-NN GoogleNet Places-365 | | k-NN ResNet50 | | MST combined desc. | |
|---|---|---|---|---|---|---|---|---|
| | bd. | - | bd. | - | bd. | - | bd. | - |
| Cor_1 | 1.0000 | 0.8638 | 0.9205 | 0.6742 | 0.9135 | 0.8242 | 1.0000 | 0.8633 |
| Cor_2 | 1.0000 | 0.5804 | 1.0000 | 0.8029 | 1.0000 | 0.7501 | 1.0000 | 0.5804 |
| Cor_3 | 0.9750 | 0.8007 | 1.0000 | 0.7369 | 0.7667 | 0.7962 | 1.0000 | 0.7857 |
| D3A | 0.6549 | 0.7832 | 0.6147 | 0.6403 | 0.5939 | 0.6906 | 0.6612 | 0.7852 |
| D7 | 0.8193 | 0.8016 | 0.8570 | 0.7277 | 0.9810 | 0.8071 | 0.8193 | 0.8041 |
| F102 | 1.0000 | 0.7144 | 0.6293 | 0.4635 | 0.9167 | 0.5764 | 1.0000 | 0.7833 |
| F104 | 0.8537 | 0.8504 | 0.4815 | 0.4471 | 0.7704 | 0.7114 | 0.8537 | 0.8632 |
| F105 | 1.0000 | 0.8813 | 0.8296 | 0.6392 | 0.9722 | 0.6380 | 1.0000 | 0.8851 |
| F107 | 0.8772 | 0.7239 | 0.2875 | 0.2526 | 0.5326 | 0.4596 | 0.8963 | 0.7224 |
| av. mAP | 0.9089 | 0.7778 | 0.7356 | 0.5983 | 0.8274 | 0.6948 | **0.9145** | 0.7859 |

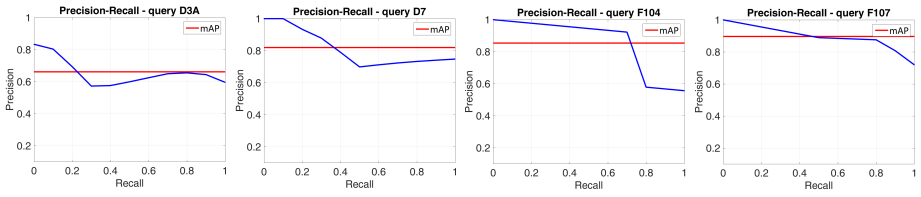


Fig. 4: Precision-recall plots for selected rooms (D3A, D7, F104, F107).

row shows some correctly matched reference images with the query image, i.e. retrieved relevant images.

Fig. 5: Query image and relevant images (upper row), irrelevant images (second row), images retrieved using NetVLAD features (images acquired in room F102).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

In the second phase of this part of the research, we performed experiments consisting in estimating the confidence of place recognition. The confidence of place recognition has been estimated as ratio of sharp images to total number of images in a pool of fifty most similar images with the image most similar to the query image, see Fig. 6a) that contains sample sub-tree. Figure 6b) illustrates the estimated confidence for all 22 landmark points. It turned out also that information about room category is important in context of confidence of robot decisions since spatial accuracy of place recognition in long and narrow corridors with similar scene content is much smaller. As we can observe, the lowest confidences are for Corridor_3 in 2nd landmark point and for Corridor_1, 1st landmark point. Experiments consisting in determining if acquired image is representative enough for scene recognition were conducted as well. For instance, if acquired image is blurry but the robot knows that it belongs to tree branch in which there are only images belonging to single room then it can decide to perform deblurring the image and then use it for place recognition.

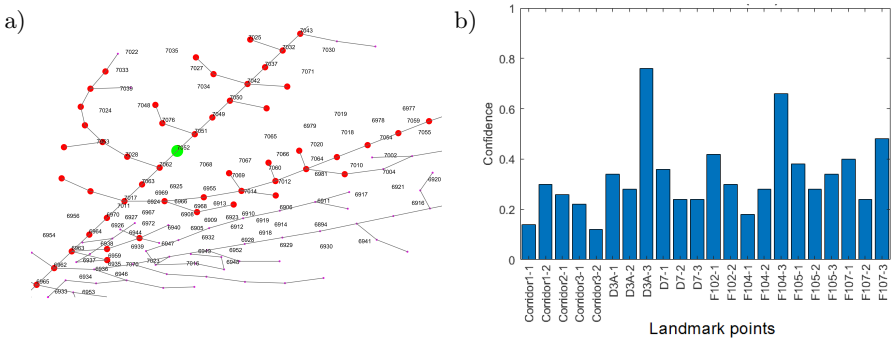


Fig. 6: Sub-tree with most similar image to the query image (green) and fifty most similar images with it (red), a), estimated confidence for all landmark points, b).

## 5 Conclusions

In this work we introduce an algorithm for blur detection on images with severe (unknown) blur. We demonstrate experimentally that the proposed algorithm outperforms recent algorithms. We propose a new algorithm which on the basis of graph-based decisions on deep embeddings and blur detections permits robust place recognition as well as delivers decision confidences. The algorithm has been evaluated on challenging dataset for visual place recognition with images acquired by a humanoid robot.

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

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

## References


1. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.: Past, present, and future of simultaneous localization and mapping: Towards the robust-perception age. IEEE Tr. on Robotics **32**(6) (2016) 1309–1332
2. Cebollada, S., Paya, L., Flores, M., Peidro, A., Reinoso, O.: A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. Expert Systems with Applications (2020) 114–195
3. Lowry, S., Snderhauf, N., Newman, P., Leonard, J., Cox, D., Corke, P., Milford, M.J.: Visual place recognition: A survey. IEEE Tr. on Rob. **32** (2016) 1–19
4. Odo, A., McKenna, S., Flynn, D., Vorstius, J.: Towards the automatic visual monitoring of electricity pylons from aerial images. In: Int. Conf. VISAPP. (2020)
5. Zhao, J., Tang, J., Zhao, D., Cao, H., Liu, X., Shen, C., Wang, C., Liu, J.: Place recognition with deep superpixel features for brain-inspired navigation. Review of Scientific Instruments **91**(12) (2020) 125110
6. Tolias, G., Avrithis, Y., Jégou, H.: Image search with selective match kernels: Aggregation across single and multiple images. Int. J. of Computer Vision **116**(3) (2015) 247–261
7. Ovalle-Magallanes, E., Aldana-Murillo, N.G., Avina-Cervantes, J.G., Ruiz-Pinales, J., Cepeda-Negrete, J., Ledesma, S.: Transfer learning for humanoid robot appearance-based localization in a visual map. IEEE Access **9** (2021) 6868–6877
8. Pretto, A., Menegatti, E., Bennewitz, M., Burgard, W., Pagello, E.: A visual odometry framework robust to motion blur. In: IEEE Int. Conf. on Robotics and Automation. (2009) 2250–2257
9. Torii, A., Arandjelović, R., Sivic, J., Okutomi, M., Pajdla, T.: 24/7 place recognition by view synthesis. In: CVPR. (2015) 1808–1817
10. Maffra, F., Chen, Z., Chli, M.: Viewpoint-tolerant place recognition combining 2D and 3D information for UAV navigation. In: IEEE Int. Conf. on Robotics and Automation (ICRA). (2018) 2542–2549
11. Garg, S., Milford, M.: Straightening sequence-search for appearance-invariant place recognition using robust motion estimation. In: Proc. of Australasian Conf. on Robotics and Automation (ACRA). (2017) 203–212
12. Chen, Z., Lam, O., Adam, J., Milford, M.: Convolutional neural network-based place recognition. In: Proc. of Australasian Conf. on Robotics and Aut. (2014) 1–8
13. Suenderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of ConvNet features for place recognition. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). (2015) 4297–4304
14. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **40**(6) (2018) 1437–1451
15. Arandjelovic, R., Zisserman, A.: All About VLAD. In: IEEE Conf. on Computer Vision and Pattern Recognition, IEEE Computer Society (2013) 1578–1585
16. Zaffar, M., Khaliq, A., Ehsan, S., Milford, M., McDonald-Maier, K.: Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions. CoRR **abs/1207.0016** (2019)
17. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **40**(6) (2018) 1452–1464
18. López-Cifuentes, A., Escudero-Viñolo, M., Bescós, J., lvaro García-Martín: Semantic-aware scene recognition. Pattern Recognition **102** (2020) 107256



ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14

19. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int J. of Computer Vision **42**(3) (2001) 145–175
20. Yandex, A.B., Lempitsky, V.: Aggregating local deep features for image retrieval. In: IEEE Int. Conf. on Computer Vision (ICCV). (2015) 1269–1277
21. Ma, J., Jiang, X., Fan, A., Jiang, J., Yan, J.: Image matching from handcrafted to deep features: A survey. Int. Journal of Computer Vision **129**(1) (2020) 23–79
22. Kwolek, B.: Visual odometry based on Gabor Filters and Sparse Bundle Adjustment. In: Proc. IEEE Int. Conf. on Robotics and Automation. (2007) 3573–3578
23. Arth, C., Pirchheim, C., Ventura, J., Schmalstieg, D., Lepetit, V.: Instant outdoor localization and SLAM initialization from 2.5d maps. IEEE Trans. on Visualization and Computer Graphics **21**(11) (2015) 1309–1318
24. Chen, Z., Maffra, F., Sa, I., Chli, M.: Only look once, mining distinctive landmarks from ConvNet for visual place recognition. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). (2017) 9–16
25. Hou, Y., Zhang, H., Zhou, S.: Evaluation of object proposals and ConvNet features for landmark-based visual place recognition. J. of Intell. & Rob. Syst. **92**(3-4) (2017) 505–520
26. Philbin, J., Isard, M., Sivic, J., Zisserman, A.: Descriptor learning for efficient retrieval. In: Computer Vision – ECCV, Springer (2010) 677–691
27. Tolias, G., Avrithis, Y., Jgou, H.: To aggregate or not to aggregate: Selective match kernels for image search. In: IEEE Int. Conf. on Comp. Vis. (2013) 1401–1408
28. Mao, J., Hu, X., He, X., Zhang, L., Wu, L., Milford, M.J.: Learning to fuse multiscale features for visual place recognition. IEEE Access **7** (2019) 5723–5735
29. Camara, L.G., Přeučil, L.: Visual place recognition by spatial matching of high-level CNN features. Robotics and Autonomous Systems **133** (2020) 103625
30. Wozniak, P., Afrisal, H., Esparza, R.G., Kwolek, B.: Scene recognition for indoor localization of mobile robots using deep CNN. In: Computer Vision and Graphics, LNCS, vol. 11114, Springer (2018) 137–147
31. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: IEEE Conf. on Computer Vision and Pattern Recognition. (2009) 413–420
32. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. In: Int. Conf. on Learning Repr. ICLR 2016. (2016)
33. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: IEEE Int. Conf. on Comp. Vis. (2003) 1470–1477
34. Zhong, C., Malinen, M., Miao, D., Fränti, P.: A fast minimum spanning tree algorithm based on k-means. Inf. Sci. **295**(C) (2015) 1–17
35. Tax, D.M.: Data description toolbox - dd tools, ver. 2.1.3. `https://github.com/DMJTax/dd_tools` (2021)
36. Narvekar, N., Karam, L.: A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). IEEE Tr. IP **20**(9) (2011) 2678–2683
37. Pech-Pacheco, J.L., Cristobal, G., Chamorro-Martinez, J., Fernandez-Valdivia, J.: Diatom autofocusing in brightfield microscopy: a comparative study. In: Proc. 15th Int. Conf. on Pattern Recognition. Volume 3. (2000) 314–317
38. Sun, J., Wenfei Cao, Zongben Xu, Ponce, J.: Learning a convolutional neural network for non-uniform motion blur removal. In: IEEE Conf. on Computer Vision and Pattern Rec. (CVPR). (2015) 769–777
39. Cun, X., Pun, C.M.: Defocus blur detection via depth distillation. In: ECCV, Springer (2020) 747–763
40. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: IEEE Conf. on Comp. Vis. and Patt. Rec. (2018) 8174–8182

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77977-1_14