Scene Recognition for Indoor Localization of Mobile Robots Using Deep CNN

Piotr Wozniak[4], Hadha Afrisal[2], Rigel Galindo Esparza[3], and Bogdan Kwolek[1]

[1] AGH University of Science and Technology, 30 Mickiewicza, 30-059 Kraków, Poland
bkw@agh.edu.pl, http://home.agh.edu.pl/ bkw/contact.html
[2] Universitas Gadjah Mada, Bulaksumur Yogyakarta 55281, Indonesia
[3] Monterrey Institute of Technology and Higher Education
Av. E. G. Sada 2501 Sur, Tecnolgico, 64849 Monterrey, Mexico
[4] Rzeszów University of Technology
Al. Powstańców Warszawy 12, 35-959 Rzeszów, Poland

Abstract. In this paper we propose a deep neural network based algorithm for indoor place recognition. It uses transfer learning to retrain VGG-F, a pretrained convolutional neural network to classify places on images acquired by a humanoid robot. The network has been trained as well as evaluated on a dataset consisting of 8000 images, which were recorded in sixteen rooms. The dataset is freely accessed from our website. We demonstrated experimentally that the proposed algorithm considerably outperforms BoW algorithms, which are frequently used in loop-closure. It also outperforms an algorithm in which features extracted by FC-6 layer of the VGG-F are classified by a linear SVM.

1 Introduction

In recent years, development of mobile robots has reached a promising milestone in terms of low-energy and effective locomotion [11, 20]. With these achievements, it is expected that in a near future, service robots will be ready to be employed in a massive scale. In many home, office or hospital settings a humanoid robot needs to navigate from one place to other places autonomously, for instance while transporting and collecting items from one room to other rooms, and so on. Any service humanoid robot is expected to work autonomously or semi-autonomously in those settings, therefore there are still many open-problems especially related to robot localization and mapping, which need to be investigated further [19].

Long-term localization and navigation in unknown environments is becoming increasingly important task for service robots. Such robots should cope with observation errors and the localization should work even in unexpected and dynamic situations, and possibly self-similar environments. The aim of Simultaneous Localization and Mapping (SLAM) is constructing or updating a map of an unknown environment while simultaneously keeping the track of the robot's location within it. Noisy measurements of the robot's odometry as well as noisy observations of the environment lead to errors that accumulate over time, growing uncertainty of each subsequent pose and map estimation. In consequence,

errors in the map estimation lead to errors in the pose estimation, and errors in pose estimation analogously lead to further errors in the map estimation [3]. In order to reliably maintain error bounds on robot's position during SLAM over long trajectories, so-called loop-closures need to be recognized, which can be achieved by collecting data associations between map and map, image and image, or images and map. Without loop-closure, the position estimated on the basis of (visual) odometry diverges from the true state since the errors accumulate over time [17]. Reliable detecting loop-closures is an important issue since any incorrect recognition of revisited places may lead to substantial mapping errors in indoor localization. Motion blur and object occlusions are among the factors that most deteriorate localization performance. Place recognition is perceived as important component towards semantic mapping and scene understanding.

The problem of qualitative robot localization refers to determining the place where the robot is present [13]. For instance, in an indoor environment a service robot should be able to determine that it is in a particular room, which may be a seminar room, lab, conference room, etc. During navigation the robot learns from experience and then recognizes previously observed places in known environment(s) and eventually categorizes previously unseen places in new rooms. This task is closely related to semantic localization, which consists in determining by a robot its location semantically with respect to objects or regions in the scene rather than reporting 6-DOF pose or position coordinates [3].

Visual place recognition algorithm is one of many solutions to provide a mobile robot with a localization ability, particularly for navigating in an indoor environment in which the robot's localization and navigation cannot only rely on the Ground Positioning System (GPS) [16]. Vision cameras are a natural choice for appearance-based SLAM, where the environment is modeled in a topological way by means of a graph. In such an approach, each graph node represents a distinctive visual location that was visited by the robot, while the edges indicate paths between places. On the basis of such a representation, the loop closure can be detected by direct image comparison, which allows us to avoid the need for maintenance and estimation of the position of the feature landmarks determined in the environment [8]. In recently proposed OpenABLE [1], the loop closure is detected on the basis of fast LDB global descriptor, which is based on idea of random comparisons between intensities of pixels in the neighborhood of the center. The main drawback of the OpenABLE is its poor scaling with the number of frames stored in the image database. Since the matching times increase linearly with the growing size of the database, the OpenABLE can only be employed in environments of limited size. Reliable recognition of the indoor place can therefore be used to constrain the searching for images only to image subset representing the currently visited room or place.

In this paper, we consider place recognition on the basis of images from a RGB camera mounted on a humanoid robot. We compare the performance of algorithms for visual recognition of the place on the basis of handcrafted features and deep learned-features. For the handcrafted features, we investigate the performance of widely used Bag-of-Words algorithm and Histogram of Uniform

Pattern (HOUP) algorithm. For the learned-features, we demonstrate the performance of transfer-learning method for VGG [26] deep neural network. We introduce a dataset for visual recognition of the indoor places, which has been recorded using Nao humanoid robot in sixteen different rooms.

2 Relevant work

There are two main approaches to achieve indoor localization of a mobile robot: SLAM and appearance-based. Visual SLAM can be computationally expensive due to the complexity connected with 3D reconstructions. Methods based on appearance can achieve good performance in coarse determining the camera location on the basis of predefined, limited set of place labels. FAB-MAP [6] is a probabilistic approach to recognize places on the basis of appearance information. The algorithm learns a generative model of the visual words using a Chow-Liu tree to model the co-occurrence probability. It performs matching the appearance of current scene to the same (similar) previously visited place through converting the images into Bag-of-Words representations built on local features such as SIFT or SURF.

Visual place description techniques can be divided into two broad categories: those that are based on local features; and those that describe the whole scene. Several local handcrafted features for visual place recognition have been investigated, such as SIFT [15] or SURF [2], binary-based local features such as BRISK [12] or ORB [21], and other feature descriptors such as lines [32], corner and edges [9], and HOUP [23]. Global or whole-image descriptors of the images, called also image signatures, such as Gist [18] process the whole image regardless of its content, and were investigated in place recognition as well [28].

One method which is commonly used in visual place recognition is Bag-of-Words (BoW) or Bag-of-Features (BoF), which usually follows three main steps: (1) extraction of local image features, (2) encoding of the local features in an image descriptor, and (3) classification of the image descriptor [4]. Sivic et al. [27] proposed an effective BoW method for object retrieval. Their method searches for and localizes all the occurrences of an object in a video, given a query image of the object. Local viewpoint invariant SIFT features are quantized on the basis of a pretrained visual word vocabulary and aggregated into a term frequency-inverse document frequency (tf-idf) vector. A benefit of the tf-idf representation is that it can be stored in the inverted file structure that yields an immediate ranking of the video frames containing the object. A successful visual place recognition technique has been achieved by Dorian's algorithm [7], which utilizes binary encoding to describe features in a very effective and efficient BoW algorithm. However, its main limitation is the use of features that lack of rotation and scale invariance. Another significant challenge in developing robust BoW-based visual place recognition algorithm is the existence of many similar and repetitive structures in man-made indoor settings, such as tiles, ceilings, windows, doors, and many more [31]. In addition, many of those methods are not robust to viewpoint and illumination changes [29].

Recent advancement in machine learning and deep learning has shed the light on a novel approach of utilizing Convolutional Neural Network (CNN) for many applications in vision-based recognition [24]. CNNs offer a new way of employing learned features for solving image retrieval problems as demonstrated in visual place recognition and categorization [5, 14]. However, implementing deep learning for mobile robots is not quite straightforward, especially to carry out complex problems such as place recognition, SLAM and pose estimation, in which a high uncertainty is a considerable challenge [30]. Another challenge is that limited battery life makes it not easy to implement and execute in real-time algorithms that are based on deep models. Additionally, there is also need to cope with blurring and rotation of images around the optical axis, which arise during locomotion of the humanoid robot.

3 Image descriptors

Below we outline descriptors that were used in our algorithm and experiments.

3.1 Handcrafted descriptors

SIFT. SIFT feature [15], called SIFT keypoint is a selected image region with an associated descriptor. The keypoints can be extracted by the SIFT detector, whereas their descriptors can be expressed by the SIFT descriptor. By searching for image blobs at multiple positions and scales, the SIFT detector provides invariance to translation, rotations, and re-scaling of the image. The SIFT descriptor is a 3D spatial histogram of the image gradients describing the content of the keypoint region. It is calculated with respect to gradient locations and orientations, which are weighted by the gradient magnitude and a Gaussian window superimposed over the region.

HOUP. Histogram of Oriented Uniform Patters is a descriptor that is calculated by filtering the image sub-block by a Gabor Filter with different orientations. The output of the Gabor filter is then used to calculate Local Binary Patterns (LBP). The Principal Component Analysis (PCA) is then executed to reduce the dimensionality of such patterns of textural features. In [23] the input image is divided into 3×3 blocks. Each block undergoes Gabor filtering, and then LBPs are computed for each block. The block has 58 uniform + 1 non uniform patterns for each orientation, and it is represented by a vector of size $59 \times 6 = 354$. The PCA reduces the dimension of each block representation from 354 to 70. The descriptor of the whole image has size equal to $70 \times 9 = 630$. Finally, Support Vector Machines (SVM) and K-nearest neighbors (k-NN) classifiers were executed to achieve place recognition.

3.2 Learned descriptors

Transfer learning is a machine learning approach in which a model developed and then learned for a certain task is reused as the starting point for a model on

another setting. It allows to reuse knowledge learned from tasks for which a lot of labeled data is available to settings, where only little labeled data is in disposal. The utilization of a pretrained models is now a common approach, particularly when patterns extracted in the original dataset are useful in the context of another setting. This is due to the reason that enormous resources are required to train deep learning models, and/or large and challenging datasets on which deep learning models can be trained. For example, the Alex-Net required about 2 - 3 weeks to train using GPU and utilized approximately 1.2 million images. It was trained on a subset of the ImageNet database, which has been utilized in ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2012)[22]. The model can classify images into 1000 object categories. It has learned rich feature representations for a wide range of images.

The pretrained deep neural networks demonstrated its usefulness in many classification tasks, including visual place recognition. However, more often in practice a technique called fine-tuning is employed. In such an approach the chosen deep model that was trained on a large dataset like the ImageNet is utilized to continue training it (i.e. running back-propagation) on the smaller dataset we have. The networks trained on a large and diverse datasets like the ImageNet capture well universal features like curves and edges in their early layers, that are useful in most of the classification tasks. Another fine-tuning technique, which can be useful if the training dataset is really small, consists in taking the output of the intermediate layer prior to the fully connected layers as the features (bottleneck features) and then learning a linear classifier (e.g. SVM) on top of it. The reason for this is that the SVMs are particularly good at determining decision boundaries on small amounts of data.

The VGG-F network is an eight layer deep convolutional neural network (DCNN), see Fig. 1, which has been originally designed and trained for image classification. Its architecture is similar to the one used by Krizhevsky et al. [10]. The input image size is $224 \times 224 \times 3$. Fast processing is ensured by the four pixel stride in the first convolutional layer. The network has been trained on ILSVRC data using gradient descent with momentum. The hyper-parameters are the same as used by Krizhevsky. Data augmentation in the form of horizontal flips, random crops, and RGB color jittering has been applied in the learning process.

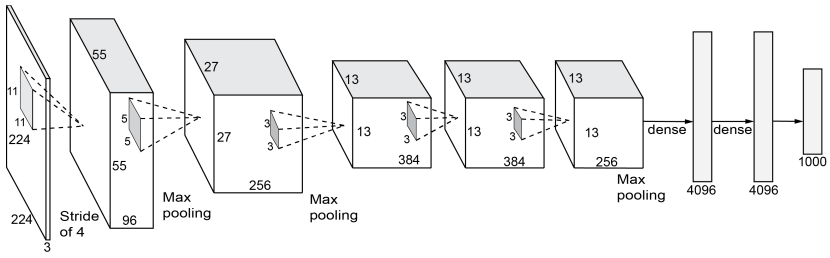

Fig. 1: Architecture of VGG-F.

The deep features can be extracted after removing the last classification layer consisting of 1000 neurons. The only image pre-processing is to resize the input images to the network input size and to subtract the average image, which is provided together with the network parameters.

4 CNN-based indoor place recognition

We investigated the features extracted by both 6th and 7th layers of VGG-F network. We compared the accuracies of place recognition using the features extracted by the layers mentioned above. The classification has been done using a linear SVM. As it turned out that far better results can be obtained on the basis of FC-6 features, the FC-6 features were used in evaluations described below. Afterwards, we removed the last layer from VGG-F and added three new layers to the layer graph: a fully connected layer, a softmax layer, and a classification output layer. The fully connected layer has been set to have the same size as the number of classes in the place recognition dataset. In order to achieve faster learning in the added layers than in the transferred layers, the learning rate factors of the fully connected layer have been increased. Such a deep neural network has been fine-tuned on place recognition datasets investigated in this work. The learning has been conducted using stochastic gradient descent with momentum (SGDM) optimizer with initial learning rate set to 0.0001, momentum set to 0.9 and L2 regularization set to 0.0001. The batch size has been set to ten samples.

The images acquired by Nao robot during walking are usually rotated around the optical axis. In general, for the CNN networks two approaches can be incorporated to encode rotation invariance: 1) employing rotations to the feature maps or alternatively to the input image, or 2) applying rotations to the convolution filters. The first approach comes down to a common practice of augmenting the training set with several rotated instances of the training images. Such a method permits the model to incorporate the rotation invariance [25]. We coped with the problem of rotated images around the optical axis, as well as scale variations through augmenting the data by image rotations in range -10,. . . ,10 deg, and shifting them horizontally and vertically in range -20,. . . ,20 pixels.

5 Experimental results

At the beginning of the experiments we conducted experimental evaluations on York Univ. Dataset [14]. The dataset has been recorded using a color camera (Point Grey Bumblebee) mounted on Pioneer and Virtual Me mobile robots in two different lighting conditions (daylight and night time). On the Pioneer robot the camera was 88 centimeters above the ground level, whereas on Virtual Me robot it was mounted 117 centimeters above the floor. The dataset consists of 29 917 images for 11 places, with 100-500 images belonging to each place. All images have been acquired with a resolution of 640×480 pixels, with the camera fixed at an upright location. During data recording the robots were manually driven at the speed of approximately 0.5 meters per second through all the eleven

places. The images were acquired at the rate of approximately three frames per second.

As in [14], we determined the accuracies of place recognition in four scenarios:

- Same Robot, Same Lighting Conditions
  Pioneer 1 – Pioneer 2 (Day - day or night - night) and Virtual Me 1 - Virtual Me
- Same Robot, Different Lighting Conditions
  Pioneer 1 – Pioneer 2 (Day - night or night - day) and Virtual Me 1 - Virtual Me 2 (Day - night or night - day)
- Different Robot, Same Lighting Conditions
  Pioneer - Virtual Me (Day - day or night - night) and Virtual Me - Pioneer (Day - day or night - night)
- Different Robot, Different Lighting Conditions
  Pioneer - Virtual Me (Day - night or night - day) and Virtual Me - Pioneer (Day - night or night - day)

The accuracy of place recognition in each scenario has been determined as the average of diagonal values in the confusion matrix.

We compared the recognition accuracies of algorithms operating both on handcrafted features and learned features. At the beginning we evaluated BoW algorithm operating on SIFT features. The classification has been performed using a k-NN as well as a linear SVM. The next algorithm was based on HOUP descriptor and a linear SVM. In the third algorithm the features extracted by a pretrained VGG-F deep neural network were classified by a linear SVM. Table 1 presents experimental results that were achieved on York Univ. Dataset [14] by mentioned above algorithms.

Table 1: Place recognition accuracy [%] on York Univ. Dataset [14], $^{[A]}$BoW using SIFT with k-NN classifier, $^{[B]}$BoW using SIFT with SVM classifier, $^{[C]}$HOUP with SVM classifier, $^{[D]}$VGG-F features classified by SVM.

Experiment	Training set	Testing set	Lighting conditions	Accuracy [%]			
				BoW+SIFT		HOUP	VGG-F, FC-6
				k-NN$^{[A]}$	SVM$^{[B]}$	SVM$^{[C]}$	SVM$^{[D]}$
1	Pioneer	Pioneer	same	68	75	98	99
2	Virtual Me	Virtual Me	same	66	77	98	98
3	Pioneer	Pioneer	different	60	72	93	94
4	Virtual Me	Virtual Me	different	62	73	93	92
5	Pioneer	Virtual Me	same	58	69	92	92
6	Virtual Me	Pioneer	same	58	68	92	95
7	Pioneer	Virtual Me	different	55	64	82	86
8	Virtual Me	Pioneer	different	58	66	85	89

As can be seen in Table 1, the algorithm based on VGG-F and the linear SVM achieves far better accuracies in comparison to results obtained in [14]. As we can observe, in experiment #7 and #8 the accuracy of VGG-F+SVM

algorithm is relatively low compared to other experiments (only 86% and 89%), i.e. when using different camera and different lighting condition for training and testing subset of images.

In the next stage of the experiments the evaluations were performed on our dataset for visual recognition of the place, which has been recorded with the Nao humanoid robot. During walking the robot acquired the images from the onboard camera. The dataset has been recorded in rooms, offices and laboratories of our department. In each of the sixteen rooms we recorded from 309 to 627 color images of size 640×480. The total number of images is equal to 8000. The dataset is available for download at http://pwozniak.kia.prz.edu.pl/ICCVG2018.html.

Table 2 presents experimental results that were achieved by the investigated algorithms. As we can observe, the SVM classifier operating on SURF features quantized by BoW achieves the lowest recognition performance. In the discussed algorithm the dataset was split in proportions 0.6, 0.2 and 0.2 for training, validation and testing parts, respectively. The C parameter of the SVM classifier has been determined experimentally in a grid search. As can been seen, the SVM operating on features extracted by FC-6 layer of the pretrained VGG-F achieves far better results. The discussed results were achieved in 10-fold cross-validation. The next row contains results that were achieved using testing data with no motion blur. The blurred images were removed from the test subset of dataset manually. We can notice, after suppression of blur noise from the test data the improvement in the recognition performance is insignificant. The best results were achieved by the fine-tuned VGG-F. The discussed results were achieved by the deep neural network that has been trained in four epochs. Having on regard that considerable part of the images is contaminated by motion blur as well as bearing in mind that many of them are rotated, we fine-tuned the VGG-F neural network on the augmented data. As can been seen in the last row of Tab. 2, the data augmentation does not improve the recognition performance.

Table 2: Place recognition performance on our dataset. $^{[A]}$ SURF, BoW, SVM classifier, $^{[B]}$ Features extracted by VGG-G, SVM classifier, $^{[C]}$ Features extracted by VGG-G, SVM classifier, no-blur, $^{[D]}$ VGG-F fine-tuned, $^{[E]}$ VGG-F fine-tuned, data augmentation.

	Accuracy	Precision	Recall	F1-score
$^{[A]}$BoW, SURF, SVM	0.7307	0.7307	0.7312	0.7310
$^{[B]}$VGG-F, SVM	0.9513	0.9510	0.9483	0.9488
$^{[C]}$VGG-F, SVM, no-blur	0.9544	0.9544	0.9549	0.9536
$^{[D]}$VGG-F fine-tuned	0.9719	0.9712	0.9720	0.9716
$^{[E]}$VGG-F fine-tuned, aug.	0.9669	0.9657	0.9676	0.9666

The experiments were conducted using scripts prepared in Matlab and Python. They were performed on a PC computer equipped in i7, 3GHz CPU with 16GB RAM and NVidia Quadro K2100M with 2GB RAM. On the GPU the classification time of single image by fine-tuned VGG-F is 0.0873 s.

6 Conclusions

In this paper we proposed an algorithm for visual place recognition on images acquired by a humanoid robot. During robot locomotion the images undergo rotations as well contamination by the motion blur. We recorded a dataset for indoor place recognition and made it publicly available. We demonstrated experimentally that a deep neural network, which has been built on the basis of the pretrained VGG-F through removing the last layer and then adding a fully connected layer, softmax layer and the output one achieves the best classification performance. We demonstrated that the learned model deals well with motion blur as well as rotations that arise during robot locomotion.

Acknowledgment. This work was supported by Polish National Science Center (NCN) under a research grant 2014/15/B/ST6/02808.

References

1. Arroyo, R., Alcantarilla, P., Bergasa, L., Romera, E.: OpenABLE: An open-source toolbox for application in life-long visual localization of autonomous vehicles. In: IEEE Int. Conf. on Intell. Transportation Systems. pp. 965–970 (2016)
2. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. European Conf. on Computer Vision **3951**, 404–417 (2006)
3. Cadena, C., Carlene, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Trans. on Robotics **32**(6), 1309–1332 (2016)
4. Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A.: The devil is in the details: An evaluation of recent feature encoding methods. British Machine Vision Conf. (BMVC) (2011)
5. Chen, Z., Lam, O., Jacobson, A., Milford, M.: Convolutional neural network-based place recognition. In: Australasian Conf. on Robotics and Automation (2014), https://eprints.qut.edu.au/79662/
6. Cummins, M., Newman, P.: FAB-MAP: Probabilistic localization and mapping in the space of appearance. Int. J. Rob. Res. **27**(6) (Jun 2008)
7. Galvez-Lopez, D., Tardos, T.: Bags of binary words for fast place recognition in image sequences. IEEE Trans. on Robotics **28**, 1188–1197 (2012)
8. Garcia-Fidalgo, E., Ortiz, A.: Vision-based topological mapping and localization by means of local invariant features and map refinement. Robotica **33**, 1446–1470 (2014)
9. Harris, C., Stephens, M.: A combined corner and edge detector. Alvey Vision Conference **15**, 10–5244 (1988)
10. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. Advances in Neural Proc. Systems pp. 1097–1105 (2012)
11. Kuindersma, S., Deits, R., Fallon, M., Valenzuela, A., Dai, H., Permenter, F., Koolen, T., Marion, P., Tedrake, R.: Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot. Advances in Neural Proc. Systems **40**, 429–455 (2016)
12. Leutenegger, S., Chli, M., Siegwart, R.: BRISK: Binary robust invariant scalable keypoints. Int. Conf. on Computer Vision (ICCV) (2011)

13. Levitt, T., Lawton, D.: Qualitative navigation for mobile robots. Artificial Intell. **44**(3), 305 – 360 (1990)
14. Li, Q., Li, K., You, X., Bu, S., Liu, Z.: Place recognition based on deep feature and adaptive weighting of similarity matrix. Neurocomputing **199**, 114–127 (2016)
15. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
16. Lowry, S., Sünderhauf, N., Newman, P., Leonard, J., Cox, D., Corke, P., Milford, M.: Visual place recognition: A survey. IEEE Trans. on Robotics **32**, 1–19 (2016)
17. Newman, P., Ho, K.: SLAM-loop closing with visually salient features. In: Proc. of IEEE Int. Conf. on Robotics and Automation. pp. 635–642 (2005)
18. Oliva, A., Torralba, A.: Building the gist of a scene: the role of global image features in recognition. In: Visual Perception, Progress in Brain Research, vol. 155, pp. 23 – 36. Elsevier (2006)
19. Oriolo, G., Paolillo, A., Rosa, L., Vendittelli, M.: Humanoid odometric localization integrating kinematic, inertial and visual information. Autonomous Robots **40**, 867–879 (2016)
20. Radford, N., Strawser, P., Hambuchen, K., Mehling, J., Verdeyen, W., Donnan, A., Holley, J., Sanchez, J., Nguyen, V., Bridgwater, L., Berka, R.: Valkryrie: NASA's first bipedal humanoid robot. J. of Field Robotics **32**, 397–419 (2015)
21. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. Int. Conf. on Computer Vision (ICCV) **32** (2011)
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. of Computer Vision **115**(3) (2015)
23. Sahdev, R., Tsotsos, J.: Indoor place recognition system for localization of mobile robots. IEEE Conf. on Computer and Robot Vision pp. 53–60 (2016)
24. Schönberger, J., Hardmeier, H., Sattler, T., Pollefeys, M.: Comparative evaluation of hand-crafted and learned local features. IEEE Conf. on Computer Vision and Pattern Recognition pp. 6959–6968 (2017)
25. Simard, P., Steinkraus, D., Platt, J.: Best practices for convolutional neural networks applied to visual document analysis. In: Int. Conf. on Document Analysis and Recognition. pp. 958–963 (2003)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
27. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering objects and their location in images. In: IEEE Int. Conf. on Computer Vision. pp. 370–377 Vol. 1 (2005)
28. Sünderhauf, N., Protzel, P.: BRIEF-Gist - closing the loop by simple means. In: IEEE/RSJ Int. Conf. on Intelligent Robots and Systems. pp. 1234–1241
29. Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M.: Place recognition with convNet landmarks: Viewpoint-robust, condition-robust, training-free. In: Proc. of Robotics: Science and Systems XII (2015)
30. Tai, L., Liu, M.: Deep-learning in mobile robotics - from perception to control systems: A survey on why and why not. arXiv (2016)
31. Torii, A., Sivic, J., Pajdla, T., Okutomi, M.: Visual place recognition with repetitive structures. Proc. of the IEEE Conf. on Comp. Vision and Pattern Recognition (2013)
32. Wang, Z., Wu, F., Hu, Z.: MSLD: A robust descriptor for line matching. Pattern Recognition **42**, 941–953 (2009)