Experimental Study of the Suitability of CNN-based Holistic Descriptors for Accurate Visual Localization

Full Paper

Alberto Jaenal, Francisco-Angel Moreno and Javier Gonzalez-Jimenez
Machine Perception & Intelligent Robotics Group (MAPIR) and Instituto de Investigacion Biomedica de Malaga (IBIMA),
University of Malaga, Teatinos Campus, 29071
Malaga, Spain
ajaenal@uma.es,famoreno@uma.es,javiergonzalez@uma.es

ABSTRACT

Holistic Image Descriptors (HIDs) are compact representations of a whole image that, being suitable for Place Recognition, are not appropriate for accurate Visual Localization. The most successful HIDs are those extracted from Convolutional Neural Networks (CNNs) like VGG, ResNet, InceptionV4 or NetVLAD. Very recently, the *equivariance* property has been proposed to reflect how image 2D transformations (e.g. rotation, flip, scale changes) influence the descriptor [17]. Our work experimentally analyzes whether such property can be a good indicator of the suitability of the existing CNN-based HID for estimating changes in the camera pose, which produces more complex transformations of the image than the pure transformations analyzed in [17]. The results we report here are a preliminary work in the context of an ongoing project towards appearance-based localization of autonomous mobile robots.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Vision for robotics**; **Image representations**;

KEYWORDS

Deep Learning, Convolutional Neural Networks, Holistic Descriptors, Equivariance, Visual Localization, Gaussian Process Particle Filters

ACM Reference format:
Alberto Jaenal, Francisco-Angel Moreno and Javier Gonzalez-Jimenez. 2019. Experimental Study of the Suitability of CNN-based Holistic Descriptors for Accurate Visual Localization. In *Proceedings of 2nd International Conference on Applications of Intelligent Systems, Las Palmas de Gran Canaria, Spain, January 7–9, 2019 (APPIS 2019)*, 6 pages.
https://doi.org/10.1145/3309772.3309800

This work has been funded by the Government of Spain under project FPU17/04512 and partially funded by the European Union under H2020-ICT projectMoveCare (732158) and by the Science and Innovation Ministry of Spain under projectWISER (DPI2017-84827-R). We gratefully acknowledge the NVIDIA GPU Grant Program for the donation of the Titan X Pascal GPU used for this research.

APPIS 2019, January 7–9, 2019, Las Palmas de Gran Canaria, Spain

ACM ISBN 978-1-4503-6085-2/19/01…$15.00
https://doi.org/10.1145/3309772.3309800

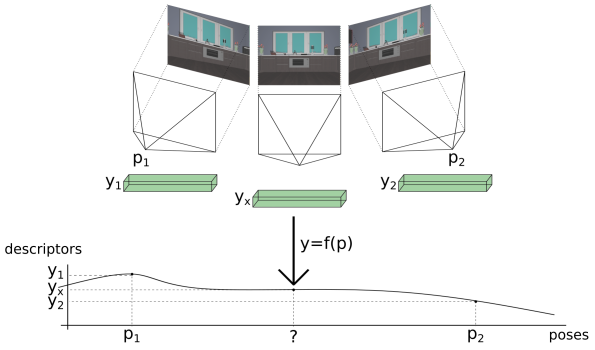

Figure 1: Representation of the function that models the relationship between poses and image descriptors. By applying regression to the visual map $\mathcal{D} = \langle (y_1, p_1), (y_2, p_2) \rangle$, we could estimate the pose of a query image descriptor $p_x = f^{-1}(y_x)$.

1 INTRODUCTION

Visual Localization (VL) is a challenging task in areas as computer vision and mobile robot navigation. It deals with the estimation of the pose (i.e. position and orientation) of a camera from the images that it captures and a previously built map. Depending on the visual features extracted from the images we distinguish between two different approaches: those based on *local* descriptors [4, 9, 21, 26], which represent geometric image features as corners, segments or blobs, and *holistic* descriptors [6, 14], which encode the *appearance* of the whole image in a compact vector.

Although local descriptors have been widely used for VL with good results in terms of accuracy, they are heavily sensitive to incorrect matches between features in the images and the map, often hindering, for example, the key process of Loop Closure in SLAM (Simultaneous Localization And Mapping). Current Holistic Image Descriptors (HIDs), on the other hand, have proven to be specially suitable for Place Recognition (a similar problem to Loop Closure) since they can be endowed with strong invariance properties to both view-point and lighting conditions. So far, however, the use of HID for VL is practically nonexistent because of the poor accuracy (being mostly limited to *Place Recognition* tasks), and completely unexplored for SLAM.

We are interested in researching what are the limitations and hurdles for that, and, eventually, in proposing algorithms for such

modality of localization, which we call *continuous appearance-based VL*. In this problem, we are given a map of the environment in the form of georeferenced image descriptors and we want to estimate the camera pose for a particular image by performing regression from the closest descriptors in the map (see Figure 1).

Concretely, in this paper we first aim at stating the particular characteristics that a descriptor should fulfil in order to achieve accurate continuous appearance-based VL, being one of them its dependency with the camera pose. This characteristic, which is contrary to typical descriptor attributes, has not been thoroughly explored in the literature but states itself as necessary for accurate appearance-based VL. Following this approach, we focus on the so-called *equivariance* property, recently proposed in [17] as an indicator of the influence of image transformations in the descriptor, and perform a comparative study of such property for a set of image descriptors generated by state-of-the-art Convolutional Neural Networks (CNNs). As the equivariance original domain of application in [17] was limited to only pure 2D transformations in the image (e.g. rotation, flip, scale change), in this paper we also contribute with the generalization of its application to arbitrary changes in the camera pose, hence enriching the study of its capacity of being used as a valid indicator for continuous appearance-based VL. Finally, those descriptors that scored best in the equivariance study were further evaluated in a localization experiment within a Gaussian Process Particle Filter (GPPF) framework, as the one presented in [19].

The preliminary results of this comparative study will be taken as a starting point for the future creation of new descriptors especially designed to encode this capability in the context of an ongoing project towards appearance-based localization of autonomous mobile robots.

2 RELATED WORKS

The most common approaches in VL work with local descriptors, since they explicitly encode the geometry of the environment, hence leading to a straightforward estimation of the camera pose through typical computer vision techniques like reprojection error minimization between the observations and the map [23, 24]. Approaches as [27, 28] exploit these features to relocalize a camera within 3d scene models, relying in 2d-to-3d feature matching for a 6 DoF accurate estimation. The main drawbacks of these methods are twofold: (i) the descriptors need to be robust and invariant to both changes in the images and in the camera Point of View (PoV), (ii) the camera pose estimation is highly dependent on the correct matching of all the features. The latter implies implementing a set of filters to avoid the presence of outliers, being specially challenging those environments with repetitive structures. Despite these, local descriptor-based systems typically achieve good results in terms of accuracy.

Alternatively, HIDs have been typically employed in classification problems such as Scene Classification [33, 34], Content-Based Image Retrieval (CBIR) [2, 14], or Place Recognition [8, 18], since they can encode the whole image with a compact, distinctive vector. This fact makes them particularly suited for Bag of Words (BoW) dictionaries [6] or classification-related techniques as Nearest Neighbours models [3] in tasks as pose retrieval.

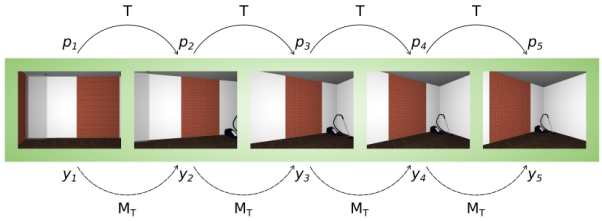

Figure 2: An equivariant descriptor y_i must satisfy that for fixed camera pose transformations T, its value must be transformed by the same map M_T.

The works in [19, 20, 22] approach appearance-based VL as a *continuous* problem. In [22] the pose of an image is interpolated between the nodes of a sequence graph, although such pose can only be estimated within the same sequence of images. Furthermore, the works in [19, 20] go further into this approach and model the descriptor as a continuous and interpolable function of the camera pose over all the space. This allows to perform regression and infer the current camera pose from the closest HIDs in the map, achieving more accurate results than typical Place Recognition systems.

Recently, the emergence of *Deep Learning* (DL) in computer vision [16] has been gradually introducing new techniques into the feature extraction field, positioning themselves as promising approaches for generating highly robust HIDs from CNNs. Although the performance of these descriptors heavily depends on a training process and the richness of the available data, they present an enormous potential in terms of environment generalization. This leads to an outstanding robustness against strong changes in the conditions in which the images were taken, including radiometric changes, dynamic objects, occlusions or even weather conditions [8], making them perfect candidates to plausibly fulfil most of the descriptor requirements for continuous VL that will be described in the section 3.

Previous works as [1, 13] linked a CNN to the camera movement, by a concept called ego-motion, which associates the pose variation to the specific image changes that they cause. *Posenet*, the approach presented in [15], regresses the camera pose by learning a whole environment through a CNN, subsequently relocalizing the camera with robustness to changes in appearance. In [3] the authors present RelocNet, a CNN whose descriptor is based on a continuous metric, using information from overlaps of the camera field-of-view. The descriptors of these CNNs, seem promising candidates for our equivariance study, but, unfortunately, they are not publicly available and, therefore, we could not include them.

3 DESCRIPTOR CHARACTERISTICS

In a nutshell, a holistic descriptor is a function that maps an image to a compact vector, but, apart from that, our approach also considers the holistic descriptor as a *continuous* and *interpolable* function of the camera pose. Thus, for a holistic descriptor to be suitable for VL, it must fulfil some essential requirements from both perspectives.

If we think about the descriptor as a function of the image, the descriptor must comply the following characteristics:

- **Invariance to radiometric changes**. A proper descriptor for VL should be robust against different illumination conditions, which are expected to change in situations in where the map is created using a particular camera (under certain illumination conditions) and later employed for localization with images taken at different times and/or by different cameras (e.g. multi-robot systems).
- **Robust response to moving objects and occlusions**. Typical localization applications will be carried out in dynamic environments in where part of the scene will remain static but the appearance of moving objects and occlusions will be frequent. Thus, the holistic descriptor must robustly generalize the localization environment and present robustness against such situations.
- **Dependence to the Point of View**. Finally, in contrast to the classification-based approaches (and also unlike local descriptors), a PoV-invariant representation would be unsuitable for continuous appearance-based VL, as they would maintain a similar value for smooth displacements, hence hindering the camera pose estimation from the descriptors stored in the map.

Although the first two characteristics are commonly desirable for HIDs (and also for local image feature descriptors), the last one opposes the tendency of what a descriptor should satisfy. Being variable as the PoV changes (or, equivalently, the camera pose) is something generally tried to be avoided in problems as Place Recognition, but desirable in continuous appearance-based VL. However, this approach has not been thoroughly explored in the literature.

On the other hand, when modelled in terms of the pose, a suitable descriptor function for continuous appearance-based VL should satisfy some attributes such as smoothness and monotony, which enhance the performance of pose regression. The formal study of the descriptor function from this perspective, though, goes beyond the scope of this paper and will be addressed in future works.

In short, we can state that, assuming that a change in pose entails a change in the image, we want the descriptor to capture such change in a proportional way but, at the same time, be invariant to changes in radiometric properties. Furthermore, the image variations caused by typical pose changes (e.g. forward movements and rotations) are usually specific and repetitive, so we state that in these cases the descriptor should vary accordingly, hence evidencing the dependence of the descriptor to the image.

In this sense, and following the studies in [17] about different mathematical properties of image representations, we focus on the *equivariance*, as it is reported to encode the relationship between image and descriptor changes under certain image transformations. In this paper, we research if such property can be employed to evaluate if a certain HID fulfils the last desirable characteristic and, therefore, it is potentially suitable to perform continuous appearance-based VL.

Thus, adopting the notation in [17], let ϕ be a function that maps an image $x \in \mathcal{X}$ to a descriptor $\phi(x) \in \mathbb{R}^d$ (with d being the descriptor length). Then, we define ϕ to be equivariant with an image transformation g if the transformation can be conveyed to ϕ. Analytically, equivariance with g is computed with the map

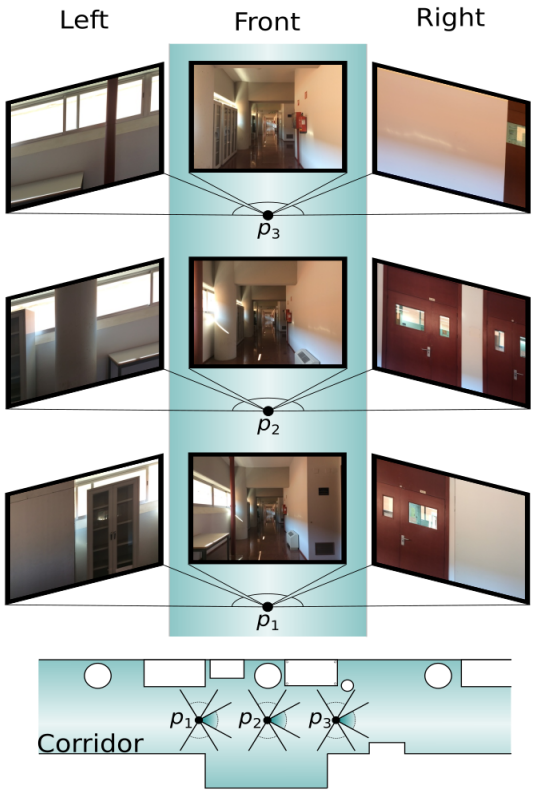

Figure 3: When navigating through corridor, the appearance of the images captured forwards is significantly more similar than the image appearance of the sides.

$M_g : \mathbb{R}^d \to \mathbb{R}^d$ such that:

$$\forall x \in \mathcal{X}: \quad \phi(gx) \approx M_g \phi(x) \tag{1}$$

It is important to note, though, that the authors of [17] limited g to be pure image transformation (e.g.: flips, rotations or affine transformations). In this work, under the assumption of pose-dependent descriptors, we generalize image transformations g to also include camera pose transformations, which cause specific changes in the image, as depicted in Figure 2.

For that, we study whether a descriptor $y_p = \phi(x_p)$ captured at camera pose p is equivariant to a certain transformation $T : \mathbb{R}^3 \to \mathbb{R}^3$. Note that, in this work, we only consider planar camera pose transformations (changes in x, y and yaw angle). Then, for any two poses as $p_{t1}, p_{t2} \in \mathbb{R}^3$ and a fixed transformation $T_{t1,\,t2}$ that relates both poses $p_{t2} = T_{t1,\,t2}\, p_{t1}$, we state that the descriptor y_p is equivariant to this transformation if there exists a map $M_{T_{t1,\,t2}} : \mathbb{R}^d \to \mathbb{R}^d$ that meets:

$$y_{p_{t2}} \approx M_{T_{t1,\,t2}} y_{p_{t1}} \tag{2}$$

However, generalizing the equivariance in this manner has a main drawback. A certain map M_T might vary for the same displacement since the appearance variation is deeply tied with the

camera orientation with respect to the scene. For example, imagine a corridor as the one shown in Figure 3. The changes in the image appearance with the movement of the camera are different when the camera aims to the end of the corridor than when the camera points to the corridor sides. The impact of this appearance changes on the descriptor must be taken into account in a proper equivariance study. Thus, our work considers translations with different camera orientations and also evaluates how the descriptor varies with camera rotations, so that we can assess the descriptor behaviour for changes in every degree of freedom of the camera movement.

4 GAUSSIAN PROCCESS PARTICLE FILTERS

In order to make this work more self-contained, in this section we briefly describe Gaussian Processes (GPs), which stands in the core of the particle filter approach employed to evaluate the performance of the chosen descriptors for continuous appearance-based VL in terms of localization accuracy.

GPs are non-parametric learning methods designed to solve probabilistic regression and classification problems. They can consistently infer the relationship between the observations (p_i, y_i) and the predictions (p_j, y_j^*) by modelling a distribution over the descriptor function. In this way, GPs can interpolate new predictions without any previous knowledge of the underlying function, using only the observations.

GPs associate the information contained in each sample by a similarity function called kernel $k(p, p')$ where, in our work, p and p' represent the poses of two HIDs. Typical approaches use kernels defined in terms of distance as the Radial Basis Function (RBF) kernel:

$$k(p, p') = \beta^2 exp(-\alpha \|p - p'\|_2^2), \tag{3}$$

where α and β are optimizable parameters.

This expression models the behaviour of the descriptor depending on the closeness of the samples and, therefore, GPs can be employed for modelling the descriptors in a continuous fashion over poses, as proposed in [19]. Based on this, they employed a GP as the observation model in a GPPF, where the GP predictions were used as the likelihood for weighting each particle.

We assume that the performance of the GP-based descriptor regression improves when working over equivariant descriptors, as they are more predictable for similar pose transformations. Consequently, we will perform GPPF localization experiments with the descriptors with the best equivariant behaviour in order to experimentally validate the equivariance study.

5 EXPERIMENTAL VALIDATION

In this section we show the evaluation of the equivariance for some CNN-based descriptors in different datasets, from which we obtained a set of images with fixed 2D movements of the camera. An example of the extracted data is shown in Figure 2, where images were taken with fixed rotations around the vertical axis (i.e. with different *yaw* angle values).

Subsequently, we have validated the results of the equivariance study by comparing the performance of different descriptors in a GPPF-based localization experiment in where a GP performs a regression over the descriptor assuming a dependence on the pose.

5.1 Study of the Descriptors Equivariance

This section focuses on the study of which CNN-based descriptors can be best modelled by the equivariance (eq. 2) for certain pose transformations. For that, since we assume the continuity of the descriptor over the pose (and the image appearance), the transformations must be constrained to ensure some image overlap.

Given that we restricted the study to 2D VL, we have evaluated the equivariance of the descriptors for fixed planar transformations. We have extracted the descriptors from several groups of images in certain poses, maintaining fixed transformations in all the three degrees of freedom.

As far as we know, there are no public approaches that specifically train descriptors to be equivariant. Consequently, we have performed the evaluation with descriptors generated by widely employed state-of-the-art CNN methods. As a metric to quantify the equivariance property, we have defined the *Normalized Accumulated Residual Score* S_T for a certain transformation T as:

$$S_T = \sum_i^N r_i^2 = \sum_i^N (y_{p_{i2}} - M_T y_{p_{i1}})^T (y_{p_{i2}} - M_T y_{p_{i1}}) \tag{4}$$

with N being the number of descriptor pairs and r_i^2 as the squared distance between the observed descriptor and the transformed one for each pair of poses $\{p_{i1}, p_{i2}\}$, given that $p_{i2} = T p_{i1}$.

We first minimized S_T through a least squares method in order to find the best M_T for each single evaluated descriptor. Then, we compared the resulting minimized score of all the descriptors in Figure 4 to find the most equivariant one.

For the evaluation, we have employed images extracted from the KITTI [7] (urban, realistic outdoor scenes), TUMIndoor [12] (realistic, indoor scenes) and SUNCG [30] (virtual, indoor scenes) datasets, since they represent a wide variety of environments, hence granting our equivariance study with a high generalization level.

Regarding the evaluated methods, the following recent CNNs have been selected:

- Image Classification CNNs: AlexNet [16], VGGNet [29], ResNet [10], Inception-v4 [31] and NasNet [35]. The feature maps of these CNN are commonly used in Transfer Learning [11] for different purposes, due to their generalization capacity.
- Scene Classification CNNs: presented in [33], trained from Transfer Learning, their feature maps can generalize better for scenes.
- Image Retrieval CNNs: CNNs from [25] and NetVLAD [2] (whose implementation in TensorFlow is given by [5]). These CNNs are purposely trained for the classification-based approach of VL.

It is important to note that the descriptors were extracted from either the latest and some intermediate layers of each CNN, performing Principal Component Analysis reduction to make them comparable.

The results of the equivariance study are shown in Figure 4, where we only depict the results of CNNs that achieved the best performance (i.e. lowest S_T). Surprisingly, Image Retrieval CNNs outperform the rest of CNNs, yielding lower residuals, even though they are supposedly PoV-invariant descriptors aimed to recognise

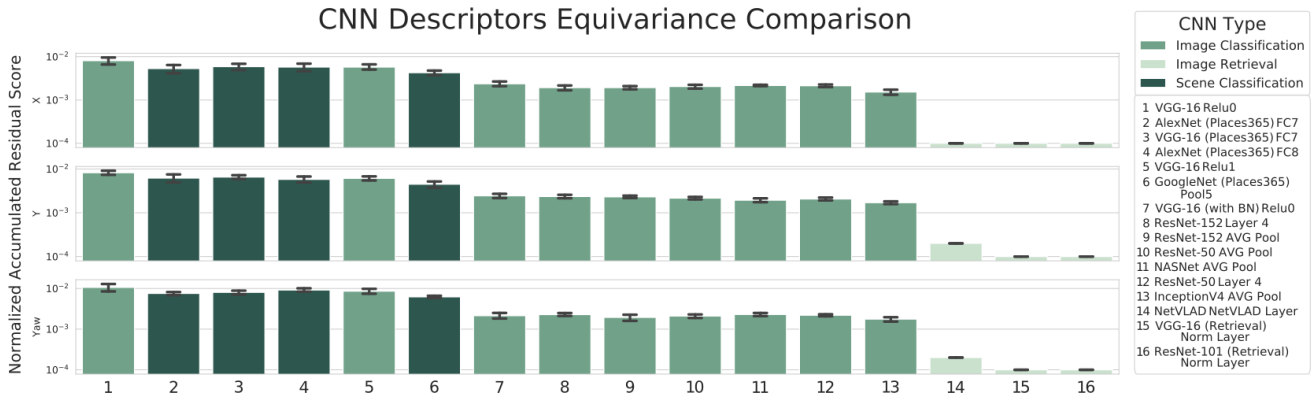

Figure 4: A comparison of the Normalized Accumulated Residual Score of different descriptors (tagged as the CNN and the layer they were extracted from) for fixed variations in the indicated degrees of freedom.

places and objects from different views. The group of Image Classification CNNs follows them in terms of accuracy, hence exhibiting their capability to generalize.

5.2 GPPF-based Localization

Finally, in order to check if the results of the equivariance study are sufficiently relevant for VL, we have compared the performance of some of the best descriptors in a GPPF-based localization experiment.

The experiments are based on the simulations presented in [19] in where a large sequence of poses and descriptors of the environment is employed as a visual map $\mathcal{D}_{\text{train}} = \langle (p_i, y_i), i = 1...N_{\text{train}} \rangle$, and a shorter sequence of descriptors $\mathcal{Q} = \langle (y_j), j = 1...N_{\text{test}} \rangle$ is used as queries to localize the camera. The camera poses corresponding to such descriptors are used as Ground Truth (GT). All these images were extracted from the SUNCG Dataset [30] with the House 3D environment [32].

The descriptor performance has been evaluated by measuring the mean error of the final position achieved by the GPPF in the described sequence in a set of 50 repetitions for each descriptor. In addition, we have repeated the localization experiment for different numbers of particles of the GPPF, in order to find a tendency on the behaviour of the descriptors in relation to such parameter.

For this evaluation, we have selected the three descriptors that best performed in the equivariance study (i.e. the nets specifically trained for Image Retrieval, NetVLAD, VGG and ResNet-101) and two descriptors from other CNNs with lower performance (VGG with Batch Normalization (BN) and NASNet), in order to confirm that less equivariant descriptors should yield worse results.

The preliminary results indicated in Figure 5 show that the equivariance is a potential indicator of the capabilities of a HID to be employed in continuous appearance-based VL within a GP-based regression approach, as those descriptors exhibiting lower S_T in equation 4 (i.e. higher equivariance) incur in lower errors in VL. The results also show that the number of particles does not reduce the error while inherently incurring in a larger computational cost, rendering increasing the number of particles worthless.

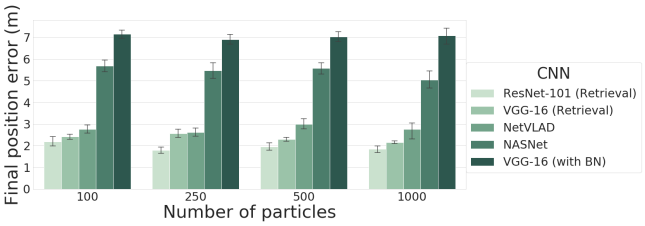

Figure 5: Mean of the final position error of the GPPF experiment with the five of the best descriptors of the equivariance study for a different number of particles.

6 CONCLUSIONS AND FUTURE WORK

In this work, we have first described the characteristics that HIDs should satisfy in order to be useful to perform continuous appearance-based VL, stating that they must be invariant to both radiometric changes and dynamic objects in the scene but also exhibit dependence to the PoV (unlike traditional, local features-based systems). Then, we have proposed the equivariance of the image as a potential indicator of the descriptor suitability for this aim, and have generalized its application to not only pure image transformation but also to changes in the camera pose.

In order to assess this, we have performed the first comparative study of HIDs generated from state-of-the-art CNN methods in terms of their equivariance. For that, we have measured the Normalized Accumulated Residual Score that appears when transforming the descriptors according to certain fixed camera 2D motions. The results indicate that descriptors created from standard Image Retrieval CNNs yield lower score (i.e. they are more equivariant) than other CNNs more related to Image or Scene Classification. Not only that, such descriptors obtain more accurate results when employed in a localization experiment with a GPPFs, hence demonstrating the equivariance potential as a measure of the descriptor VL suitability.

However, these preliminary results should be considered in the context of an ongoing project towards appearance-based VL of

autonomous mobile robots. Thus, as future works, we envisage to address the following:

- Further research about the properties and limitations of the equivariance parameter for VL.
- The creation of CNN-based descriptors especially trained to be equivariant, in order to achieve better results in regression.
- The development of techniques to create an optimal visual map with an equivariant descriptor, achieving good performance with the minimum number of samples.
- The design of specific kernel functions for an optimum GP-based regression of equivariant descriptors.

REFERENCES

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. 2015. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*. 37–45.

[2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5297–5307.

[3] Vassileios Balntas, Shuda Li, and Victor Prisacariu. 2018. RelocNet: Continuous Metric Learning Relocalisation using Neural Nets. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 751–767.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *European conference on computer vision*. Springer, 404–417.

[5] Titus Cieslewski, Siddharth Choudhary, and Davide Scaramuzza. 2018. Data-efficient decentralized visual SLAM. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2466–2473.

[6] Mark Cummins and Paul Newman. 2008. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research* 27, 6 (2008), 647–665.

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Ruben Gomez-Ojeda, Manuel Lopez-Antequera, Nicolai Petkov, and Javier González Jiménez. 2015. Training a Convolutional Neural Network for Appearance-Invariant Place Recognition. *CoRR* abs/1505.07428 (2015). arXiv:1505.07428 http://arxiv.org/abs/1505.07428

[9] Ruben Gomez-Ojeda, David Zuñiga-Noël, Francisco-Angel Moreno, Davide Scaramuzza, and Javier Gonzalez-Jimenez. 2017. PL-SLAM: a Stereo SLAM System through the Combination of Points and Line Segments. *arXiv preprint arXiv:1705.09479* (2017).

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[11] Minyoung Huh, Pulkit Agrawal, and Alexei A Efros. 2016. What makes ImageNet good for transfer learning? *arXiv preprint arXiv:1608.08614* (2016).

[12] Robert Huitl, Georg Schroth, Sebastian Hilsenbeck, Florian Schweiger, and Eckehard Steinbach. 2012. TUMindoor: An extensive image and point cloud dataset for visual indoor localization and mapping. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 1773–1776.

[13] Dinesh Jayaraman and Kristen Grauman. 2015. Learning image representations tied to ego-motion. In *Proceedings of the IEEE International Conference on Computer Vision*. 1413–1421.

[14] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. 2012. Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence* 34, 9 (2012), 1704–1716.

[15] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*. 2938–2946.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[17] Karel Lenc and Andrea Vedaldi. 2015. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 991–999.

[18] Manuel Lopez-Antequera, Ruben Gomez-Ojeda, Nicolai Petkov, and Javier Gonzalez-Jimenez. 2017. Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognition Letters* 92 (2017), 89–95.

[19] Manuel Lopez-Antequera, Nicolai Petkov, and Javier Gonzalez-Jimenez. 2016. Image-based localization using Gaussian processes. In *Indoor Positioning and Indoor Navigation (IPIN), 2016 International Conference on*. Winner, Best Paper award.

[20] Manuel Lopez-Antequera, Nicolai Petkov, and Javier Gonzalez-Jimenez. 2017. City-scale continuous visual localization. In *Mobile Robots (ECMR), 2017 European Conference on*. IEEE, 1–6.

[21] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, Vol. 2. Ieee, 1150–1157.

[22] Will Maddern, Michael Milford, and Gordon Wyeth. 2012. CAT-SLAM: probabilistic localisation and mapping using a continuous appearance-based trajectory. *The International Journal of Robotics Research* 31, 4 (2012), 429–451.

[23] Francisco-Angel Moreno, Jose-Luis Blanco, and Javier Gonzalez-Jimenez. 2016. A constant-time SLAM back-end in the continuum between global mapping and submapping: application to visual stereo SLAM. *The International Journal of Robotics Research* 35, 9 (2016), 1036–1056.

[24] Raúl Mur-Artal and Juan D. Tardós. 2017. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *IEEE Transactions on Robotics* 33, 5 (2017), 1255–1262. https://doi.org/10.1109/TRO.2017.2705103

[25] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. 2017. Fine-tuning CNN Image Retrieval with No Human Annotation. *CoRR* abs/1711.02512 (2017). arXiv:1711.02512 http://arxiv.org/abs/1711.02512

[26] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *Computer Vision (ICCV), 2011 IEEE international conference on*. IEEE, 2564–2571.

[27] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. 2011. Fast image-based localization using direct 2d-to-3d matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 667–674.

[28] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. 2018. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proc. CVPR*, Vol. 1.

[29] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). arXiv:1409.1556 http://arxiv.org/abs/1409.1556

[30] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. 2017. Semantic Scene Completion from a Single Depth Image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition* (2017).

[31] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *CoRR* abs/1602.07261 (2016). arXiv:1602.07261 http://arxiv.org/abs/1602.07261

[32] Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018. Building Generalizable Agents with a Realistic and Rich 3D Environment. *arXiv preprint arXiv:1801.02209* (2018).

[33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[34] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[35] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2017. Learning Transferable Architectures for Scalable Image Recognition. *CoRR* abs/1707.07012 (2017). arXiv:1707.07012 http://arxiv.org/abs/1707.07012