An Evaluation between Global Appearance Descriptors based on Analytic Methods and Deep Learning Techniques for Localization in Autonomous Mobile Robots

Sergio Cebollada[1], Luis Payá[1], David Valiente[1], Xiaoyi Jiang[2] and Oscar Reinoso[1]

[1]*Department of Systems Engineering and Automation, Miguel Hernández University, Elche, 03202, Spain*

[2]*Department of Computer Science, University of Münster, Münster, 48149, Germany*

Keywords: Mobile Robots, Omnidirectional Images, Global Appearance Descriptors, Localization, Deep Learning.

Abstract: In this work, different global appearance descriptors are evaluated to carry out the localization task, which is a crucial skill for autonomous mobile robots. The unique information source used to solve this issue is an omnidirectional camera. Afterwards, the images captured are processed to obtain global appearance descriptors. The position of the robots is estimated by comparing the descriptors contained in the visual model and the descriptor calculated for the test image. The descriptors evaluated are based on (1) analytic methods (HOG and *gist*) and (2) deep learning techniques (auto-encoders and Convolutional Neural Networks). The localization is tested with a panoramic dataset which provides indoor environments under real operating conditions. The results show that deep learning based descriptors can be also an interesting solution to carry out visual localization tasks.

1 INTRODUCTION

Nowadays, the use of visual information to solve mobile autonomous robotic tasks is widely expanded. In these cases, the robot must be able to build a map within the environment and estimate its position within that environment. These tasks are known as mapping and localization. Among the different sensors used, the omnidirectional cameras introduce an interesting solution since they are able to provide information that covers a field of view of 360 *deg*.

Global appearance descriptors have been proposed by several authors to extract characteristic information from images and use this information for mapping and localization. For instance Zhou *et al.* (Zhou et al., 2018) propose the use of the descriptor *gist* to solve the localization through matching the best keyframe in the dataset based on the given robot's current view. Korrapati and Mezouar (Korrapati and Mezouar, 2017) introduced the use of omnidirectional images through global appearance descriptors to create a topological mapping approach and a loop closure detection method. More recently, Román *et al.* (Román et al., 2018) evaluate the use of global appearance descriptors for localization under illumination changes. In this work, several distance measurements were also evaluated with the aim to obtain a similitude distance between images which represents the geometrical distance between the positions where those images were captured.

The computation of these descriptors is based on analytic methods, nevertheless, during the last years, some authors have proposed the use of deep learning techniques to create global appearance descriptors. For example, on the one hand, Xu *et al.* (Xu et al., 2016) proposed the use of auto-encoders to detect histopathological images of breast cancer. On the other hand, Xu *et al.* (Xu et al., 2019) used a CNN-based descriptor to obtain the most probable position within an indoor map through Monte Carlo Localization and also to solve the kidnapping problem; Payá *et al.* (Payá et al., 2018) proposed also the use of the CNN-based descriptors but in this case for hierarchical mapping. Both works are based on the net *places* (Zhou et al., 2014). The descriptors extracted from this network correspond to the ones calculated in some of the fully convolutional layers within the network.

Through this work, we carry out a comparison between global appearance descriptors based on analytic methods and global appearance descriptors based on deep learning techniques to solve the visual localization task. The goodness of these methods are measured according to the accuracy (error of localiza-

Cebollada, S., Payá, L., Valiente, D., Jiang, X. and Reinoso, O.
An Evaluation between Global Appearance Descriptors based on Analytic Methods and Deep Learning Techniques for Localization in Autonomous Mobile Robots.
DOI: 10.5220/0007837102840291
In *Proceedings of the 16th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2019)*, pages 284-291
ISBN: 978-989-758-380-3

tion) and computing time (to calculate the descriptor and to estimate the position of the robot).

The remainder of the paper is structured as follows: Section 2 explains the algorithm used to estimate the position of the test images within the environment. After that, section 3 outlines the global appearance descriptors which will be evaluated. Section 4 explains and presents the dataset used as well as the experimental results and the discussion about them. Finally, section 5 outlines the conclusions and future research lines.

2 LOCALIZATION METHOD

The localization task consists in an image retrieval problem. This is, obtaining the image which presents higher similitude in relation to the new captured image. For this purpose, the robot has previously obtained visual information from the environment, i.e., global appearance descriptors that are calculated from the N_{Train} images captured from different positions of the environment. This task is known as mapping and this step must be carried out before starting the localization. Therefore, the localization task is solved through the following steps:

- The robot captures a new omnidirectional image from an unknown position.
- That image is transformed to panoramic (im_{test}) and after that, the corresponding descriptor is calculated ($\vec{d}_{test}$).
- Once the descriptor is available, the robot calculates the cosine distance (selected in (Cebollada et al., 2019) as the best distance method for global appearance descriptors) between the test descriptor ($\vec{d}_{test}$) and each descriptor from the visual model ($\vec{d}_j$, where $j = 1, .., N_{Train}$).
- A vector of distances is obtained as $\vec{h}_t = \{h_{t1}, ..., h_{tN_{Train}}\}$ where $h_{tj} = dist\{\vec{d}_{test}, \vec{d}_j\}$.
- The node which presents the minimum distance($d_t^{nn} | t = argmin_j h_{tj}$) corresponds to the estimated position of the robot.

3 THE GLOBAL APPEARANCE DESCRIPTORS

Visual localization has been commonly solved either using local features along a set of scenes or using a unique descriptor per image which contains information on its global appearance. These second methods are know as global appearance description and have been used to solve the localization task since they allow straightforward localization algorithms. For instance, Naseer *et al.* (Naseer et al., 2018) propose a localization method from global appearance (by using histogram of oriented gradients descriptors and features from deep convolutional neural networks) to solve the localization problem and to keep in parallel several possible trajectories hypotheses.

Basically, the steps to calculate a global appearance descriptor are the following: (1) The starting point is a panoramic image expressed as a bidirectional matrix ($im_j \in \mathbb{R}^{M_x \times M_y}$). (2) Then the specific mathematical calculations are applied and a vector which characterizes the original image will be obtained ($\vec{d}_j \in \mathbb{R}^{l \times 1}$ and corresponds to the image im_j).

The first global appearance descriptors used in computer vision were descriptors based on analytic methods. Nevertheless, during the last years, the emergence of the deep learning techniques have empowered the use of descriptors based on these new methods.

3.1 Methods based on Analytic Methods

These methods are basically based on calculations of gradients and orientation of the different pixels which compose the image. Their use has been quite often to solve mobile robotics issues. For instance, Su *et al.* (Su et al., 2017) used a global descriptor to reduce pose search space with the aim to solve the kidnapped robot problem in indoor environments under different conditions. An interesting study was carried out by Román *et al.* (Román et al., 2018) which evaluates the use of global appearance descriptors for localization under illumination changes. More recently, Cebollada *et al.* (Cebollada et al., 2019) evaluate the use of global appearance descriptors to build hierarchical maps through clustering algorithms and then to solve the localization in those maps.

Among the different methods, this work proposes the use of HOG and gist, which have been used in previous works (Cebollada et al., 2019) and have proved to present interesting results for localization tasks.

Regarding the **HOG** descriptor, it was introduced by Dalal and Triggs (Dalal and Triggs, 2005) to solve the detection of pederastians. In this work, the procedure is the one proposed by Leonardis and Bischof (Leonardis and Bischof, 2000): the panoramic image is divided into k_1 horizontal cells and a histogram of gradient orientation (with b bins per histogram) is compiled per each cell. Finally, the set of histograms are arranged in a unique row to compose the final descriptor $\vec{d} \in \mathbb{R}^{b \cdot k_1 \times 1}$

As for the ***gist*** descriptor, it was introduced by Oliva *et al.* (Oliva and Torralba, 2006). In this work, the version used consists on: (1) obtaining m_2 different resolution images, (2) applying Gabor filters over the m_2 images with m_1 different orientations, (3) grouping the pixels of each image into k_2 horizontal blocks and (4) arranging the obtained orientation information into one row to create a vector $\vec{d} \in \mathbb{R}^{m_1 \cdot m_2 \cdot k_2 \times 1}$.

3.2 Methods based on Deep Learning

During the last years, the use of deep learning methods to solve computer vision issues has extensively grown. Regarding the localization task through the use of visual information, this work studies the use of Convolutional Neural Networks (CNN) and the use of auto-encoders. The idea is to obtain vectors which characterize the images through some deep learning technique. On the one hand, these methods can result very interesting since their use can be focused on specific kind of images (such as indoor environments in our case) and, hence, providing more efficient descriptors. On the other hand, these methods lead to previous training which normally implies huge processing data and noteworthy time.

Regarding the use of **CNNs**, these networks have been commonly designed for classification. In this sense, (1) a set of images correctly labeled are collected and introduced into the network to tackle the learning process and after that, (2) the network is properly available to face the classification (test image as input and the CNN outputs the most likely label option). The CNNs are composed by several hidden layers whose parameters and weights are tuned through the training iterations. In this work, some hidden layers outputs are used to obtain global appearance descriptors. This idea have already been proposed by some authors such as Mancini *et al.* (Mancini et al., 2017), who use them to carry out place categorization with the Naïve Bayes classifier or Payá *et al.* (Payá et al., 2018), who proposed CNN-based descriptors to create hierarchical visual models for mobile robot localization. The CNN architecture that has been used in this work is *places* (Zhou et al., 2014), which was trained with around 2.5 million images to categorize 205 possible kinds of scenes (no re-training is carried out in this work). Fig. 1 shows the architecture of the *places* CNN, which is based on the caffe CNN. The net basically consists in (1) an input layer, (2) several intermediate hidden layers and (3) an output layer. Within the intermediate layers, the first phase consists in (2.1) layers for featuring learning (whose layers incorporate several filters and the output generated are used as input for the next layer) and (2.2) layers for classification (whose layers are fully connected and they generate vectors which provide information for classification).

In this work, we have evaluated the output information from 5 layers. Three fully convolutional layers ('*fc6*', '*fc7*' and '*fc8*') whose output size are 4096×1, 4096×1 and 205×1 respectively. Moreover, we have obtained two descriptors from the output of 2D convolution layers ('*conv4*' and '*conv5*'). These layers apply several sliding convolutional filters to the input images with the aim to activate certain characteristics of the image. Hence, the output of these layers is a set of images which are the input image after being filtered. Finally, a descriptor is basically obtained from these layers through selecting an image from the output dataset and arranging the data (matrix) in a single row (vector). Since the size of the output images is 13×13, the size of the descriptor is 169×1.

As for the use of **auto-encoders**, the aim of these neural networks is to reconstruct the output through compressing the input into a latent-space representation (Hubens, 2018). The fig. 2 shows the architecture design of the auto-encoders. These networks firstly compress the input (encoding) and secondly reconstruct the input departing from the latent space representation (decoding). The idea consists in building a latent representation to obtain useful features with small dimension, i.e., training the auto-encoder to extract the most salient features. For example, Gao and Zhang (Gao and Zhang, 2017) used auto-encoders to detect loops for visual Simultaneous Localization And Mapping (SLAM).

For this experiment, two types of auto-encoder are proposed. Both have been trained using the same parameters (Coefficient for the L_2 weight regularizer, 0,004; Coefficient that controls the impact of the sparsity regularizer, 4; Desired proportion of training examples a neuron reacts to, 0.15; Encoder Transfer Function, "Logistic sigmoid function"; and Maximum number of training epochs, 1000) and also both have been trained using a GPU (NVIDIA GeForce GTX 1080 Ti), but whereas the first option (*auto-enc-Frib*) is trained with the images obtained from the dataset used to evaluate the localization (explained in sec. 4), the second alternative (*auto-enc-SUN*) is trained with images obtained from a dataset (SUN 360 DB (Xiao et al., 2012)) which contains generic panoramic images. The aim of this second option is to create a generic auto-encoder based on indoor panoramic images which provides a good-enough solution to obtain descriptors for panoramic images independently the environment. This solution would solve the handicap that introduces the descrip-

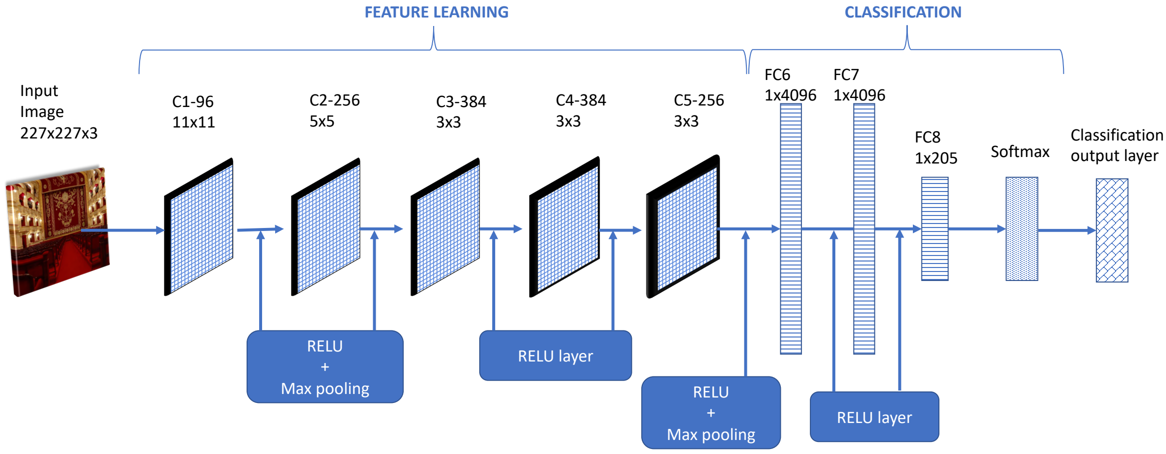

Figure 1: CNN architecture design of the pre-trained 'caffe' model.

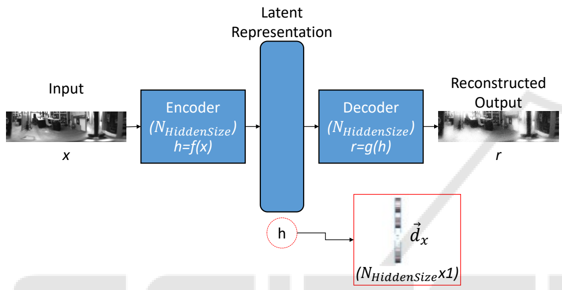

Figure 2: Auto-encoder architecture design and extraction of features departing from the latent representation.

tor based on auto-encoders regarding the need to carry out a previous training before calculating the descriptors. For *auto-enc-Frib*, the training dataset consists in 519 panoramic images whose size is 512 × 128; for *auto-enc-SUN*, the training dataset consists in 541 panoramic images whose size is also 512 × 128. Furthermore, the auto-encoders are trained varying the size of hidden representation of the auto-encoder (this number is the number of neurons in the hidden layer) and the resultant descriptor size obtained depends directly on that number ($N_{HiddenSize} \times 1$). Regarding the computing time to train these auto-encoders, the computer needs between 6 min and 2,94 hours, being directly proportional to the number of neurons (the more number of neurons there are, the more computing time is required).

4 EXPERIMENTS

4.1 Dataset

The experiments were carried out through the use of the COLD dataset (Pronobis and Caputo, 2009), which contains visual information along a trajectory. It contains three indoor laboratory environments in three cities (Freiburg, Saarbrücken and Ljubljana) and three different illumination conditions. Nevertheless, for the experiments purposes, only the images related to the Freiburg environment were used and no illumination changes have been considered, i.e., the images used were only captured under cloudy conditions (during the light hours but the sunlight does not considerably affect the shots). This lack of illumination changes is due to the fact that this work is focused on studying the goodness of the descriptors for localization task, however, in future works, an extension to study the illumination changes effects will be considered. This dataset includes changes in the environment such as people walking or position of furniture and objects. An example of these dynamic conditions can be seen in fig. 3.

Figure 3: Panoramic image from COLD database.

Among the different paths, the red one was selected for this experiment because it is the longest. Afterwards, the images are split into two datasets: training and test datasets. Training dataset is composed by 519 images which present an average distance around 20 cm between an image and the following one. The test dataset is composed by 2595 images and the average distance between images is 4,10 cm. The table 1 shows the information about the datasets in detail.

Table 1: Number of images in each room of the training and test datasets created from the Freiburg environment.

Name	**Number of images in Training**	**Number of images in Test**
Printer area	44	223
Corridor	212	1044
Kitchen	51	255
Large Office	34	175
2-persons office 1	46	232
2-persons office 2	26	131
1-person office	31	154
Bathroom	49	247
Stairs area	26	134
Total number	519	2595

4.2 Evaluation of the Localization

To evaluate the goodness of each descriptor method for localization, two parameters are considered: On the one hand, (1) the average localization error, which measures the Euclidean distance between the position estimated and the real position where the test image was captured. To obtain this value, the ground truth provided by the dataset is used. Nevertheless, the ground truth is only used for this purpose (it is not to solve the localization task). On the other hand, (2) the average computing time, which is analyzed through two values, (2.a) the computing time to calculate the descriptor and (2.b) the computing time to estimate the position of the test image.

The results obtained through the use of analytic descriptors (HOG and *gist*) and the descriptors based on deep learning (auto-encoders and CNNs) are shown in the tables 2, 3 and 4. These tables show the size of the descriptor, the average localization error (cm), the average computing time to calculate the descriptor (ms) and the average computing time to estimate the position of the test images (ms).

Regarding the results obtained through the use of descriptors based on analytic methods (see table 2), for the HOG case, the localization error does not significantly decrease as the size increases; the computing time to calculate the descriptor is also barely constant but the time to estimate the pose increases as the size of the descriptor does. Hence, the descriptor whose size is 64 is considered the best option, because this configuration presents good accuracy and the minimum computing time. Regarding the *gist* descriptor, the localization error decreases millimetres as the size of the descriptor increases, however the time to calculate the descriptors as well as the time to estimate the pose increases significantly as the size does. Therefore, in this case, the minimum size is selected as the best option.

As for the descriptors obtained through the use of auto-encoders (see table 3), for both cases (auto-enc-Frib and auto-enc-SUN), the outputs obtained by using the auto-encoders whose size of hidden representation (number of neurons) is 10 show the worst localization error results. In the case of auto-enc-Frib, the descriptors obtained from auto-encoders with $N_{HiddenSize} = 50-500$ behaves well (localization error between 7,04 and 7,45 cm), but for the auto-enc-SUN, only the case $N_{HiddenSize} = 500$ outputs similar values. Regarding the computing times (to compute the descriptor and to estimate the pose), the longer the size of the descriptor is, the more time the method needs. Furthermore, the computing time values increase severally as the size does. For instance, in the case of $N_{HiddenSize} = 500$, with auto-enc-Frib and auto-enco-SUN, the average time are 1166 ms and 1125 ms respectively. Therefore, for auto-enc-Frib, the best configuration is reached through the auto-encoder whose number of neurons is 100, because the localization error is the minimum and the computing time is the third lowest. For auto-enc-SUN, despite the configuration with $N_{HiddenSize} = 500$ presents the worst times, it is selected as the best one because the rest of options do not provide solutions that can be used to solve the localization task.

Finally, for the CNN-based descriptors case (see table 4), in general, all the layers evaluated present good results. The first layers achieve an accuracy of around 5 cm. This behaviour is reasonable since the aim of the first layers in a CNN is to obtain global characteristic information from the images and the further CNN layers are focused on optimizing the classification task. Special consideration for the layers *'conv4'* and *'conv5'*, whose use to obtain global appearance descriptors is scarce until today and they present very optimal solutions. Regarding the computation time to calculate the descriptor, none of the layers need high values and, as it was expected, the further the corresponding layer is, the higher the time is. Moreover, the computing time to estimate the pose is directly proportional to the size of the descriptor, but the layer *'conv5'* needs less time than *'conv4'*. Hence, *'conv5'* is selected as the best layer to calculate descriptors.

5 CONCLUSIONS

In this work, a study is tackled regarding the use of global appearance descriptors for localization. This task is solved as an image retrieval problem. A dy-

Table 2: Results obtained through the use of global appearance descriptors based on analytic methods (HOG and *gist*) to solve visual localization.

Descriptor	**Size**	**Error loc. (cm)**	**Time comp. descriptor (ms)**	**Time pose est. (ms)**
HOG	**64**	$16,34 \pm 0,78$	$44,64$	$0,38$
	128	$16,23 \pm 0,73$	$45,27$	$0,51$
	256	$16,22 \pm 0,69$	$45,33$	$2,48$
	512	$16,17 \pm 0,69$	$46,52$	$4,75$
gist	**128**	$5,19 \pm 0,18$	$10,30$	$0,45$
	256	$5,11 \pm 0,17$	$11,98$	$2,19$
	512	$5,09 \pm 0,16$	$21,21$	$4,17$
	1024	$5,08 \pm 0,16$	$40,07$	$10,72$

Table 3: Results obtained through the use of global appearance descriptors based on auto-encoders (auto-enc-Frib and auto-enc-SUN) to solve visual localization.

Descriptor	**Size**	**Error loc. (cm)**	**Time comp. descriptor (ms)**	**Time pose est. (ms)**
auto-enc-Frib	10	$599,83 \pm 3,83$	49,79	$0,25$
	50	$8,61 \pm 2,29$	$138,64$	$0,44$
	100	$7,04 \pm 0,85$	249,55	$0,59$
	200	$7,45 \pm 0,23$	473,59	$0,93$
	500	$7,22 \pm 0,19$	1166,49	$4,54$
auto-enc-SUN	10	$362,73 \pm 22,77$	54,99	0,28
	50	$520,85 \pm 29,66$	$138,61$	$0,43$
	100	$916,16 \pm 31,58$	$252,39$	$0,59$
	200	$327,25 \pm 21,39$	$477,48$	$0,90$
	500	$5,31 \pm 0,34$	$1125,06$	$4,66$

Table 4: Results obtained through the use of global appearance descriptors based on *places* CNN (layers *'conv4'*, *'conv5'*, *'fc6'*, *'fc7'* and *'fc8'*) to solve visual localization.

Layer	**Size**	**Error loc. (cm)**	**Time comp. descriptor (ms)**	**Time pose est. (ms)**
conv4	169	$5,03 \pm 0,02$	$6,64$	$1,62$
conv5	**169**	$5,09 \pm 0,17$	$6,66$	$0,63$
fc6	4096	$5,14 \pm 0,18$	$7,42$	$34,38$
fc7	4096	$16,71 \pm 0,84$	$8,58$	$33,22$
fc8	205	$24,22 \pm 6,44$	$8,88$	$0,72$

namic dataset with panoramic images has been used to evaluate the experiments. Five global appearance descriptors have been evaluated: two based on analytic methods (HOG and *gist*), two based on auto-encoders and one based on CNN layers. The size of each descriptor is varied through either tuning some parameters (such as the number of bins in HOG or the size of hidden representation of the auto-encoders) or selecting a different layer in the CNN case. The localization error, the computing time to calculate the descriptor and the computing time to estimate the position of the robot have been used as parameters to measure the efficiency of these descriptors. The fig. 4 shows the results obtained for the best configuration of each descriptor evaluated. From that figure, we can conclude that the minimum localization error is obtained through the CNN-based descriptor option, but the *gist* descriptor and the auto-enc-SUN descriptor show results quite similar. The CNN-based descriptor introduces also the best option regarding the computing time to calculate the descriptor. Nevertheless, regarding the time to estimate the pose of the robot, HOG is the fastest.

Regarding the use of auto-encoders, using an auto-encoder which has been trained with images that belong to the environment outputs good-enough accuracy results. The general auto-encoder proposed through training a generic panoramic dataset works acceptably in the case of high size of hidden representation, hence this leads to high computing times. Nevertheless, its use as tool to obtain global appearance descriptors for panoramic images would be valid and

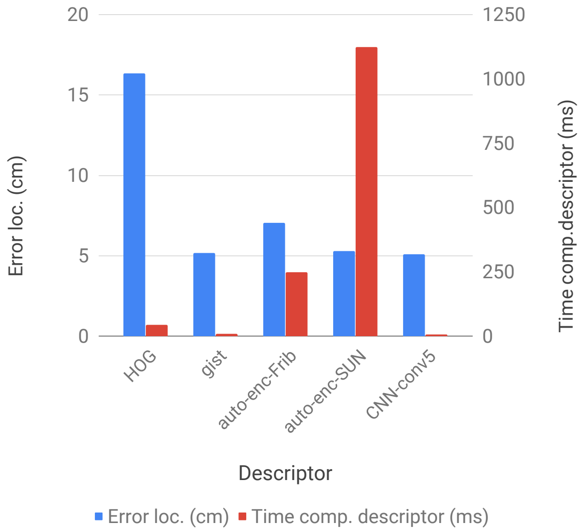

Figure 4: Summary of the best configuration for each descriptor studied.

the advantage of this method is that the auto-encoder is trained just once, then the tool is suitable independently the environment.

As for the use of CNN-based descriptors, we have proved that the first layers can output very interesting descriptors despite these are not fully convolutional layers (typically proposed to obtain descriptors). Moreover, the descriptors related to the *'conv4'* and *'conv5'* layers have produced the optimal localization solutions among all the methods evaluated: size of descriptor relatively small (which leads to fast times to estimate the position), low computing time to calculate the descriptor and very accurate localization (average error around 5 cm for a test dataset and a training dataset whose average distance between images is around 4 cm and 20 cm respectively).

ACKNOWLEDGEMENTS

This work has been supported by the Generalitat Valenciana through grant ACIF/2017/146 and by the Spanish government through the project DPI 2016-78361-R (AEI/FEDER, UE): “Creación de mapas mediante métodos de apariencia visual para la navegación de robots.”

REFERENCES

Cebollada, S., Payá, L., Mayol, W., and Reinoso, O. (2019). Evaluation of clustering methods in compression of topological models and visual place recognition using global appearance descriptors. *Applied Sciences*, 9(3):377.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients fot human detection. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, USA. Vol. II, pp. 886-893.*

Gao, X. and Zhang, T. (2017). Unsupervised learning to detect loops using deep neural networks for visual slam system. *Autonomous robots*, 41(1):1–18.

Hubens, N. (2018). *Deep inside: Autoencoders.* https://towardsdatascience.com/deep-inside-autoencoders-7e41f319999f. Accessed February 11, 2019.

Korrapati, H. and Mezouar, Y. (2017). Multi-resolution map building and loop closure with omnidirectional images. *Autonomous Robots*, 41(4):967–987.

Leonardis, A. and Bischof, H. (2000). Robust recognition using eigenimages. *Computer Vision and Image Understanding*, 78(1):99–118.

Mancini, M., Bulò, S. R., Ricci, E., and Caputo, B. (2017). Learning deep nbnn representations for robust place categorization. *IEEE Robotics and Automation Letters*, 2(3):1794–1801.

Naseer, T., Burgard, W., and Stachniss, C. (2018). Robust visual localization across seasons. *IEEE Transactions on Robotics*, 34(2):289–302.

Oliva, A. and Torralba, A. (2006). Building the gist of ascene: the role of global image features in recognition. In *Progress in Brain Reasearch: Special Issue on Visual Perception.Vol. 155.*

Payá, L., Peidró, A., Amorós, F., Valiente, D., and Reinoso, O. (2018). Modeling environments hierarchically with omnidirectional imaging and global-appearance descriptors. *Remote Sensing*, 10(4):522.

Pronobis, A. and Caputo, B. (2009). COLD: COsy Localization Database. *The International Journal of Robotics Research (IJRR)*, 28(5):588–594.

Román, V., Payáá, L., and Reinoso, O. (2018). Evaluating the robustness of global appearance descriptors in a visual localization task, under changing lighting conditions. In *ICINCO 2018. 15th International Conference on Informatics in Control, Automation and Robotics, Porto (Portugal), 29-31 July 2018*, pages 258–265.

Su, Z., Zhou, X., Cheng, T., Zhang, H., Xu, B., and Chen, W. (2017). Global localization of a mobile robot using lidar and visual features. In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 2377–2383.

Xiao, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2012). Recognizing scene viewpoint using panoramic place representation. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2695–2702.

Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., and Madabhushi, A. (2016). Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, 35(1):119–130.

Xu, S., Chou, W., and Dong, H. (2019). A robust indoor localization system integrating visual localization aided by cnn-based image retrieval with monte carlo localization. *Sensors*, 19(2):249.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495.

Zhou, X., Su, Z., Huang, D., Zhang, H., Cheng, T., and Wu, J. (2018). Robust global localization by using global visual features and range finders data. In *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 218–223.