A Deep Learning Tool to Solve Localization in Mobile Autonomous Robotics

Sergio Cebollada[a], Luis Payá[b], María Flores[c], Vicente Román[d], Adrián Peidró[e] and Oscar Reinoso[f]

Department of Systems Engineering and Automation, Miguel Hernández University, Elche, Spain

Keywords: Mobile Robotics, Omnidirectional Images, Holistic Description, Deep Learning, Hierarchical Localization.

Abstract: In this work, a deep learning tool is developed and evaluated to carry out the visual localization task for mobile autonomous robotics. Through deep learning, a convolutional neural network (CNN) is trained with the aim of estimating the room where an image has been captured, within an indoor environment. This CNN is not only used as tool to solve a room estimation, but it is also used to obtain global-appearance descriptors of the input image from its intermediate layers. The localization task is addressed in two different ways: globally, as an image retrieval problem and hierarchically. About the global localization, the position of the robot is estimated by using a nearest neighbour search between the holistic description obtained from a test image and the training dataset (using the CNN to obtain the descriptors). Regarding the hierarchical localization method, first, the CNN is used to solve the rough localization step and after that, it is also used to obtain global-appearance descriptors; second, the robot estimates its position within the selected room through a nearest neighbour search by comparing the obtained holistic descriptor with the visual model contained in that room. Throughout this work, the localization methods are tested with a visual dataset that provides omnidirectional images from indoor environments under real-operation conditions. The results show that the proposed deep learning tool is an efficient solution to carry out visual localization tasks.

1 INTRODUCTION

Over the past few years, omnidirectional imaging has been proposed by several authors to solve mobile autonomous robotics tasks, since it has proved to be a robust option (Payá et al., 2017). This type of cameras are able to provide a high quantity of information from the environment that surrounds them, with a field of view of 360 deg., with only one snapshot. For instance, Abadi *et al.* propose an omnidirectional vision system to detect obstacles through an algorithm to carry out autonomous navigation (Abadi et al., 2015). More recently, Liu *et al.* propose accurate estimation of the position and orientation of the robot within an outdoor environment by means of omnidirectional images (Liu et al., 2018).

To carry out the localization and mapping tasks by using this visual information, an extraction of the most relevant information must be tackled. Among the two most common methods, this work proposes the use of global-appearance (or holistic) description methods, since this methodology leads to more direct localization algorithms based on a pairwise comparison between descriptors. For example, Korrapati and Mezouar use global-appearance descriptors to create topological maps by means of omnidirectional images (Korrapati and Mezouar, 2017), and Do *et al.* use global-appearance description together with Group LASSO Regression to develop an autonomous mobile navigation (Do et al., 2018).

As for the hierarchical localization, the process conducted in previous works such as (Pronobis and Jensfelt, 2011) or (Payá et al., 2018) consists basically in (1) carrying out a rough but fast localization in a high-level map composed by representative descriptors and after that, (2) solving the fine localization in a low-level map composed by the instances that are represented by the descriptors selected in the rough step. These previous works have proved the effectiveness

[a] https://orcid.org/0000-0003-4047-3841
[b] https://orcid.org/0000-0002-3045-4316
[c] https://orcid.org/0000-0003-1117-0868
[d] https://orcid.org/0000-0002-3706-8725
[e] https://orcid.org/0000-0002-4565-496X
[f] https://orcid.org/0000-0002-1065-8944

Cebollada, S., Payá, L., Flores, M., Román, V., Peidró, A. and Reinoso, O.
A Deep Learning Tool to Solve Localization in Mobile Autonomous Robotics.
DOI: 10.5220/0009767102320241
In *Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2020)*, pages 232-241
ISBN: 978-989-758-442-8

of hierarchical maps to solve the localization problem departing from global-appearance descriptors obtained from omnidirectional images. In particular, the aim of the present work is to carry out the localization task in indoor environments using omnidirectional visual information as a simple image retrieval problem and also to solve the localization by means of hierarchical topological models.

Regarding the use of Artificial Intelligence (AI), these techniques have been proposed in many contributions to improve the performance of mapping and localization algorithms in mobile robotics. For instance, Dymczyk *et al.* propose the use of a classifier to classify landmark observations and conduct the localization task more robustly (Dymczyk et al., 2018). Meattini *et al.* present a human-robot interface in which the robot learns the optimal hand configuration for grasping through electromyography sensors and merging pattern recognition and factorization techniques (Meattini et al., 2018). Within the AI approaches, the deep learning branch has gained much popularity in solving these problems by means of computer vision. These methods try to construct automatically high level data models through architectures that allow linear, non-linear, multiple and iterative transformations (Bengio et al., 2013) from the initial data matrices. The idea is to train the architecture to reach a model that is capable of creating representations which best define the inputs. Regarding the robotics topic, a number of previous works propose the use of deep learning techniques. For example, Lenz *et al.* use a deep learning approach to solve the problem of detecting robotic grasps (Lenz et al., 2015); as for mobile robotics, Zhu *et al.* propose deep reinforcement learning to address target-driven visual navigation (Zhu et al., 2017). The aim of the present work is to solve the visual localization task through Convolutional Neural Networks (CNNs), since these networks have been successfully used to solve computer vision applications such as face recognition or navigation in self-driving cars. The idea in this case is to create a CNN that is able to distinguish between different rooms from an indoor environment in order to estimate correctly in which room the robot currently is. There are well known CNN architectures, such as AlexNet, that was introduced by Krizhevsky *et al.* (Krizhevsky et al., 2012). This network consists of eight layers (five convolutional layers and three fully connected layers) with a final 1000-way softmax and three pooling layers, and it is trained to classify images into 1000 object categories. GoogLeNet was proposed by Szegedy *et al.* (Szegedy et al., 2015). It has 22 layers and it is also trained for object classification but it uses 12 times fewer parameters than AlexNet. A broad review of the more outstanding CNNs can be found in (Pak and Kim, 2017).

These popular networks, together with many others that have produced successful results, have been used in the present work as starting point to develop new tools with different objectives, that is, these CNNs are reused to carry out different tasks. We use the following methods to adapt these networks to our needs.

- Reusing **common CNN architectures**. Transfer learning is a technique that consists in reusing the architecture and parameters of a CNN as a starting point to build a new CNN with a different aim. The main idea is to get profit of most of the intermediate layers, because their parameters have been tuned from millions of images and contain useful information. This technique can save a huge amount of time for training and even obtain better results than creating a new network from scratch. This idea has already been used by authors such as Wozniak *et al.*, who use the transfer learning technique to retrain the VGG-F network to classify places among 16 rooms acquired by a humanoid robot (Wozniak et al., 2018). Nevertheless, transfer learning works only if no early layers need to be modified, because the downstream architecture and parameters are no longer valid. Therefore, in these situations, transfer learning can not be used and training a network from scratch is necessary. Creating an entire network architecture is complex, hence, rather than trying to build an architecture from scratch, the present work proposes to develop the CNN through using common architectures developed by experts. In this way, the approach is similar to transfer learning (starting with pre-existing architectures), but starting from scratch with the parameters tuning.
- Generation of global-appearance descriptors from the **intermediate layers activation**. The process is basically the following. Once the CNN is properly available to face the desired task, the hidden layers perform vector description which originally is used to solve the CNN task, but it can be extracted as a global-appearance descriptor of the input image and used for a different purpose. This idea has already been proposed by some authors such as Mancini *et al.*, who use this visual information to carry out place categorization with a Naïve Bayes classifier (Mancini et al., 2017). Payá *et al.* propose CNN-based descriptors to create hierarchical visual models for mobile robot localization (Payá et al., 2018). Moreover, Cebollada *et al.* (Cebollada et al., 2019) tackle an evaluation of global-appearance descriptors obtained

from different layers of the pre-trained *places* CNN (Zhou et al., 2014) for mobile localization.

Therefore, the objective of this work is to evaluate the performance of convolutional neural networks which have been adapted and used to carry out the mapping and localization tasks for mobile robotics in indoor environments. The proposed experiments will measure the efficiency of this tool through its ability to estimate the position of the robot and the computing time required for it. Additionally, only images obtained by an omnidirectional vision system are used as source of information to solve the mapping and localization tasks. These images are obtained from an indoor dataset captured under real-operation conditions.

The remainder of the paper is structured as follows. Section 2 presents briefly the CNN developed for this work. Section 3 explains the localization method proposed by means of the deep learning tool. After that, section 4 outlines the experiments that were carried out to evaluate the validity of the proposed method for localization. Last, section 5 presents the conclusions and future works.

2 THE CONVOLUTIONAL NEURAL NETWORK DEVELOPED

As section 1 outlines, the objective of this work is to develop and test a localization framework which performs efficiently in mobile robotics through visual information. A CNN is proposed as tool to carry out this task. The aim is to solve the visual localization hierarchically. This paper presents the idea of developing a CNN which is able to estimate the room in which the robot captured the image. Afterwards, a holistic descriptor is obtained from an intermediate layer of the same CNN to estimate more accurately the position of the image within the predicted room. This process will be explained deeply in section 3. Hence, a classification CNN must be developed firstly, to estimate the room within the environment.

The CNN basically consists in predicting the label of the given input data (in this case, images). The labels (also known as targets) represent the possible categories within the environment. Before using this tool for prediction, the model requires a training with a huge variety of input data (x_{train}) and their corresponding labels (y_{train}). Then, the CNN is ready to receive new data (x_{test}) and estimate their categories ($y_{estimated}$).

2.1 The Dataset

The dataset of images used to train the CNN is the Freiburg Dataset, which has been obtained from the COLD (COsy Localization Database) database (Pronobis and Caputo, 2009). The COLD database is composed by images captured from different indoor environments through several sensors under three illumination conditions (cloudy days, sunny days and at nights) and they are also affected by presence of dynamic changes such as people walking or furniture changes and also by the blur effect. These images were captured following a trajectory along the whole environment. Among all the images provided, this work uses the omnidirectional images captured from the Freiburg environment. This dataset is also used to evaluate the localization task. Nevertheless, before training the CNN, a conversion from omnidirectional to panoramic images is carried out with the aim of comparing the obtained results with other global-appearance description methods based on panoramic or standard images. Additionally, the use of panoramic images constitutes an interesting option, since CNNs traditionally work with conventional (non panoramic) images.

Fig. 1 shows the bird's eye view of the Freiburg environment and the path that the robot traversed to obtain the images. The images of the Freiburg dataset were captured in 9 different rooms. The cloudy dataset was captured during cloudy days and it is the least affected by illumination conditions. Hence, this dataset is used as training dataset. The sunny and night datasets provided by the Freiburg COLD DB are used to evaluate the localization task under changes of illumination. Additionally, in order to establish a trustworthy comparison with previous works, the dataset is downsampled with the objective of obtaining visual information with a distance of 20 cm between consecutive images. The resulting images compose the training dataset and the rest of images are used to create a test dataset which will be used to evaluate the CNN accuracy and also the efficiency of the hierarchical localization method proposed. Table 1 shows the datasets used for this work and the number of images that each of them contains.

Due to the amount of parameters which compose a CNN, a large image dataset is required to tune them. Nevertheless, the datasets available to solve a specific task are not always as large as required to train a CNN from scratch and then, the deep model trained can not reach enough accuracy. This issue has been commonly solved through data augmentation. This technique basically consists in creating new data by applying different effects over the original images. To cite

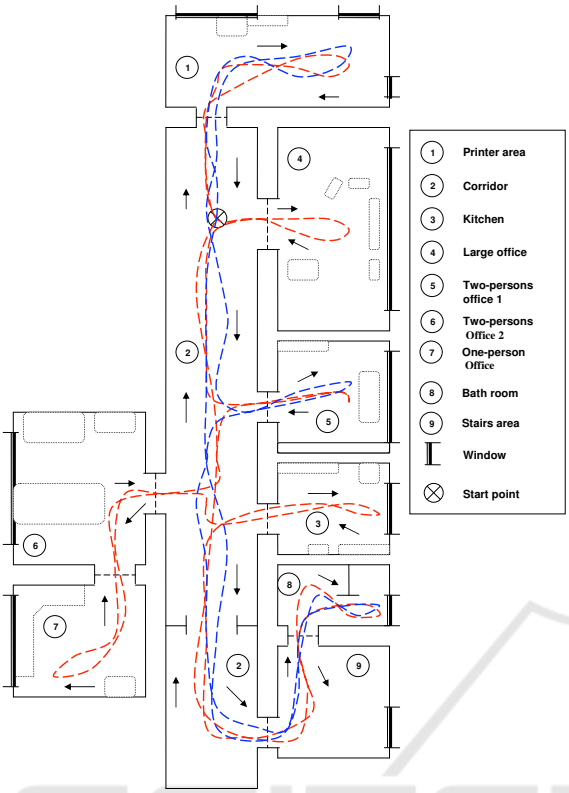


Figure 1: Bird's eye view of the Freiburg and environment. Extracted from (Ullah et al., 2007). The red dashed line is the path selected to obtain the images.

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

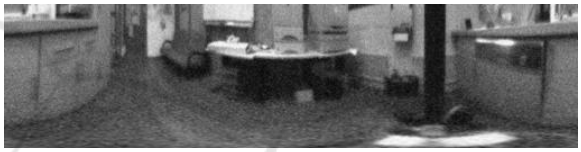
Name	Number of images in Training	Number of images in Test
1. Printer area	44	285
2. Corridor	212	1182
3. Kitchen	51	229
4. Large Office	34	132
5. 2-persons office 1	46	233
6. 2-persons office 2	26	158
7. 1-person office	31	218
8. Bathroom	49	190
9. Stairs area	26	151
Total number	519	2778

one example, Guo and Gould proposed to use data augmentation to improve a CNN training with the aim of solving an object detection task (Guo and Gould, 2015). The data augmentation proposed in this work consists in applying visual effects over the original images that can actually occur when images are captured in real-operation conditions: Random rotation,

(a) Original.

(b) Rotated.

(c) Dark.



(d) Noise.

Figure 2: Example of data augmentation. (a) Original image captured within the Freiburg environment. An effect is applied over each image: (b) random rotation, (c) darkness, (d) Gaussian noise.

reflection, darkness/brightness addition to the image, Gaussian noise, occlusions and blur effect. The fig. 2 shows examples of some of the effects applied over an original image. Hence, in order to train the CNN, instead of using the 519 images of the original training dataset, the network is trained with the augmented version (composed by 49824 images).

2.2 The Architecture and Training

In this work, we propose to use the AlexNet architecture as the base of the proposed CNN tool. The choice of AlexNet as starting point architecture to carry out the learning is due to the successful performance showed by other authors regarding its use for transfer learning such as (Han et al., 2018) and also for the simplicity of its architecture.

Therefore, some layers of the AlexNet architecture are replaced to adapt the output to the classification task desired (estimation among the 9 rooms which belong to the Freiburg environment in this work) and also to receive panoramic images as input. As for the replacement of layer to achieve the

classification desired, the three last layers which are replaced are the fully connected layer fc_8, the softmax layer and the classification layer. Additionally, regarding the input layer, since this layer was configured in AlexNet to receive $227 \times 227 \times 3$ images, it is replaced to receive $128 \times 512 \times 3$ images. Through this last change, despite the parameters of the convolutional layers are reset, we avoid a resizing of the input images which could affect their resolution and hence, effectiveness of the network created. After these changes of the original CNN, the network is ready to be trained with the new data in the training dataset. The fig. 3 shows the final architecture used along this work. We trained the CNN off-line on NVIDIA GEFORCE GTX 1080TI ®GPU system. The training time was around 4 hours. After every 30 iterations, the performance of the partially trained network was evaluated by using the data for validation.

3 MAPPING AND LOCALIZATION THROUGH THE CNN

As explained previously in section 1, one of the aims of this work is to use the holistic descriptors generated by the intermediate layers of the CNN to carry out the localization task. Regarding this description method, it basically consists in introducing the image into the CNN and retaining the data stored in one of the layers. In the case of the fully connected layers which compose the classification phase, they directly provide data arranged in a vector, hence, these data can be directly used as global-appearance descriptor. Apart from the descriptors obtained from the fully connected layers, Cebollada *et al.* showed that the data from the 2D convolutional layers $conv_4$ and $conv_5$ in the training phase are also interesting to obtain characteristic information from the images (Cebollada et al., 2019). For these layers, the data are arranged in N_{ch} matrices, where N_{ch} is the number of channels in the convolutional layer. Hence, first, a channel is selected among the rest, and after that, the matrix is re-arranged in a vector that is used as descriptor. The descriptors obtained from the $conv_4$ and $conv_5$ layers led to better localization results than the descriptors obtained from fc_6, fc_7 and fc_8. This is due to the fact that CNNs learn to detect features like color and edges in the first convolution stages and then, in deeper layers, the network learns more complicated features related to the problem to solve (in the case of AlexNet, object classification). Moreover, the size of the descriptors obtained from the convolutional layers are smaller, hence, the localization algorithm requires lower computing time.

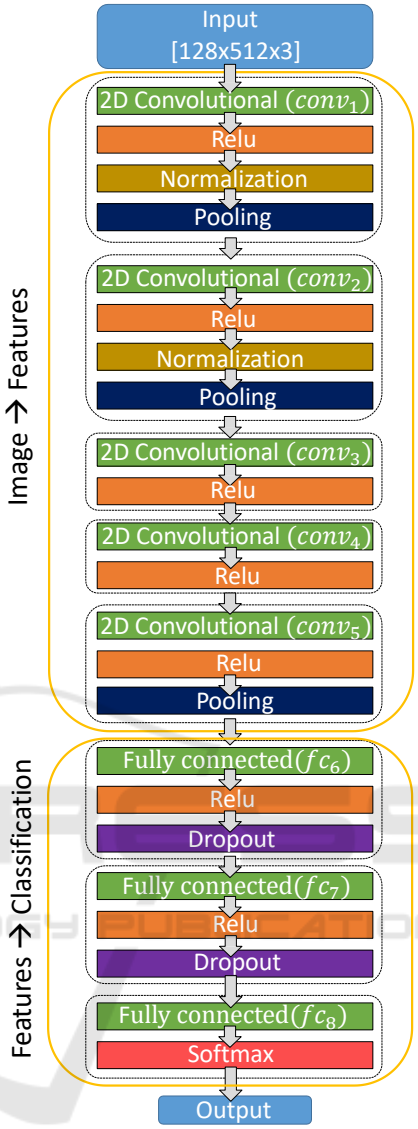


Figure 3: The CNN architecture created by departing from the AlexNet architecture. The input layer is replaced to receive images with $[128 \times 512 \times 3]$ size and the last three layers (fc_8, softmax and the classification layer) are also replaced to adapt the network to the classification task proposed.

Therefore, this work evaluates the use of the layers $conv_4$, $conv_5$, fc_6, fc_7 and fc_8 of the retrained CNN to obtain holistic descriptors for solving the mapping and localization tasks. This paper also presents a comparison between these global-appearance descriptors and classic descriptors based on analytic tools such as HOG (Histogram of Oriented Gradients) (Dalal and Triggs, 2005) or *gist* (Oliva and Torralba, 2006) to solve the mapping and localization

task by means of panoramic images.

This way, regarding the mapping task, the CNN is trained for two purposes: (1) estimating the room (used in the first step of the hierarchical localization) and (2) obtaining holistic description information from a layer (used to solve the conventional localization task as an image retrieval problem and also in the second step of the hierarchical localization).

As for the conventional localization task, the whole process is as follows. A test image im_{test} is captured from an unknown position within the environment. The holistic descriptor $\vec{d}_{test}$ is obtained from the CNN and after that, it is compared with all the descriptors contained in the training dataset $D = \{\vec{d}_1, \vec{d}_2, ..., \vec{d}_{Ntrain}\}$ and the most similar descriptor $\vec{d}_k$ is retained. Last, the position of im_{test} is estimated as the coordinates where im_k was captured.

Finally, if a hierarchical localization alternative is desired instead of the conventional method, the process conducted in previous works such as (Payá et al., 2018) or (Cebollada et al., 2019) consists basically in an nearest neighbour search with different levels of granularity. Nevertheless, the process proposed in this work consists in the following. A test image im_{test} is introduced into the CNN and an estimation about the most likely room c_i in which the image was captured is tackled (rough localization step). Apart from the estimated room, the CNN also provides the holistic descriptor $\vec{d}_{test}$ from a selected layer. Afterwards, a nearest neighbour search is carried out (fine localization step). That is, the obtained descriptor $\vec{d}_{test}$ is compared with the descriptors $D_{c_i} = \{\vec{d}_{c_i,1}, \vec{d}_{c_i,2}, ..., \vec{d}_{c_i,N_i}\}$ from the training dataset which belong to the predicted room c_i, and then, the most similar descriptor $\vec{d}_{c_i,k}$ is retained. Finally, the position of im_{test} is estimated as the coordinates where $im_{c_i,k}$ was captured. The fig. 4 shows a diagram regarding the hierarchical localization method proposed in the present work.

4 EXPERIMENTS

The training of the CNN, as well as the experiments detailed in this section have been carried out with a PC with a CPU Intel Core i7-7700 ® at 3.6 GHz. Moreover, the training of the CNN was tackled with a GPU NVIDIA GEFORCE GTX 1080TI ®. This paper presents two experiments. Additionally, the datasets presented in the subsection 2.1 were used to carry out the training of the CNN, the mapping task and later evaluation of the localization method proposed.

Throughout the experiments tackled to evaluate the goodness of the localization methods, two parameters are considered to check the accuracy and efficiency: (1) the average localization error, which measures the Euclidean distance between the position estimated and the real position where the test image was captured (obtained by the ground truth); and (2) the average computing time required to estimate the position of the test image.

4.1 Experiment 1: Comparison between Localization Methods

This subsection presents the results obtained with the proposed localization algorithm, which uses the global-appearance descriptors obtained from different layers of the trained CNN. Moreover, these results are also compared with other global-appearance description methods based on classical analytic methods, whose configuration is selected from previous works (Cebollada et al., 2019). The results obtained through the use of analytic descriptors (HOG and *gist*) and the descriptors based on deep learning are shown in the table 2. This table shows the size of the descriptor, the average localization error (cm) and the average computing time to estimate the position of the test images (ms). Regarding the localization error, the descriptor obtained from the layer $conv_4$ presents the minimum value, followed by the descriptors from the layers $conv_5$ and fc_6. As for the computing time, the fastest option is also achieved with the $conv_4$ layer, since the data obtained for this layer are calculated in a very early stage of the CNN architecture and the holistic descriptor calculated from this layer has a relatively small size. In general, the values obtained through using the CNN trained with the Freiburg training dataset to obtain the holistic descriptors improves the localization task in comparison to the descriptors calculated by analytical methods. Considering the localization error and the computing time measured, either layers $conv_4$, $conv_5$, fc_6 or fc_7 can be considered to carry out this task. Nevertheless, despite fc_8 outputs good computing time, the localization error obtained is quite worse comparing to the rest of layers. The information provided by this descriptor with 9 components allows fast computations but the information provided is not enough to characterize the main information of the images.

4.2 Experiment 2: Hierarchical Localization

As it was explained in section 3, the hierarchical localization consists in solving the localization task in

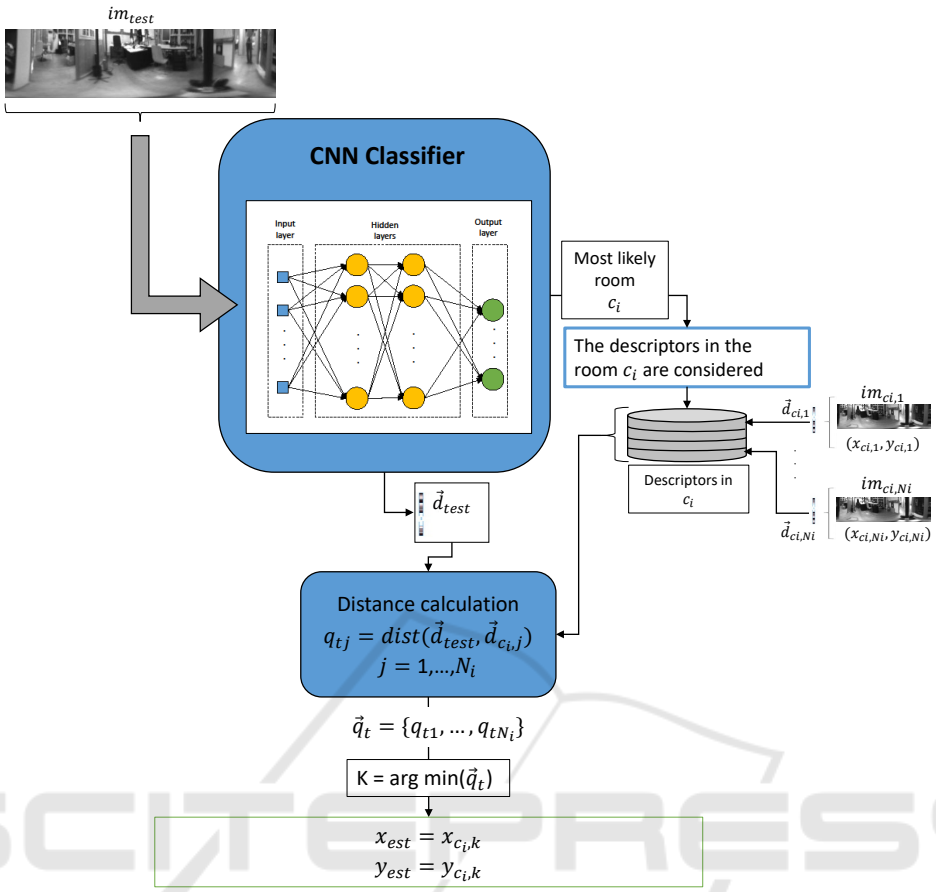


Figure 4: Hierarchical localization diagram. After capturing a test image im_{test}, it is introduced into the CNN and the most likely room is estimated c_i. At the same time, the holistic descriptor $\vec{d}_{test}$ is obtained from one of the layers and a nearest neighbour search is done with the descriptors from the training dataset included in the room predicted. The most similar descriptor (the one which produces the minimum distance with $\vec{d}_{test}$) is retained. The position of im_{test} is estimated as the position where $im_{c_i,k}$ was captured.

Table 2: Conventional localization results obtained through the use of the holistic descriptors obtained from the Freiburg CNN and through the use of *gist* and HOG description methods. The table shows the size of descriptor, the average localization error and the average computing time.

	Descriptor	Size	Avg. Error (cm)	Avg. Computing time (ms)
Layers from Freiburg CNN	$conv_4$	180	5.07 ± 0.17	6.7
	$conv_5$	180	5.09 ± 0.17	7.7
	fc_6	4096	5.09 ± 0.17	44.55
	fc_7	4096	5.14 ± 0.18	46.26
	fc_8	9	16.60 ± 29.72	7.52
Analytical methods	*gist*	128	5.19 ± 0.18	10.75
	HOG	64	16.34 ± 0.78	45.02

several steps. The hierarchical localization proposed through this work is based on two steps: a rough and a fine localization. As for the rough localization step, an evaluation of the trained CNN is carried out. To train the network, the performance basically consists in using the pre-trained CNN together with the augmented dataset by following training options. The obtained CNN is evaluated with the cloudy test dataset by introducing these images into the network and obtaining the percentage of accuracy $acc_{\%}$, which is calculated as $acc_{\%} = (N_{ok}/N_{test}) \times 100$, where N_{ok} is the number of images whose room is correctly predicted and N_{test} is the total number of images that compose the cloudy test dataset. Through this evaluation, the accuracy obtained is 98.71%. Additionally, the fig. 5 shows the confusion matrix obtained. From it, we can observe that few wrong predictions are produced. Furthermore, all these mistakes are produced

with wrong rooms which are adjacent to the correct one. For instance, in the case of the images that belong to the stairs area and were wrongly classified, the mistaken rooms were the bathroom and the corridor, which are contiguous to the correct room. Therefore, the conclusion is that the trained CNN is ready to predict in which room the input image was captured.

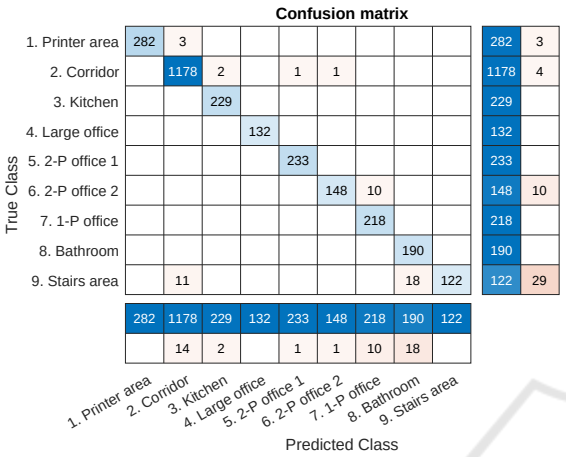


Figure 5: Confusion matrix obtained after solving the first step of the hierarchical localization (room classification) with all the test images, with the trained CNN.

Regarding the fine localization step, it consists in finding the nearest neighbour by comparing the holistic descriptor obtained from a layer of the CNN and the descriptors of the trained dataset included in the predicted room. This experiment has evaluated the efficiency of the five holistic descriptors obtained from the different CNN layers by means of measuring the average error and the average computing time calculated according to the process described in fig. 4 to carry out the localization task. Moreover, with the aim of comparing the results obtained through the proposed method with other hierarchical localization methods, this experiment also establishes a comparison between the proposed method (rough step with CNN and fine step with nearest neighbour) and the methods evaluated in previous works (Cebollada et al., 2019) (rough and fine steps solved by nearest neighbour). For these methods of previous works, the global-appearance descriptors used are the descriptor *gist* and the descriptor obtained from the layer fc_6 of the AlexNet; and the high-level map is composed by 10 representatives which were selected by using a spectral clustering algorithm. Fig. 6 shows the results obtained through using the method proposed with the different descriptors from the CNN layers and also the results obtained by the methods proposed in previous works.

Regarding the different methods evaluated to carry out the hierarchical localization, this experiment shows that the method proposed performs substantially better than alternative methods previously proposed. Fig. 6 shows that the five description methods based on the Freiburg CNN present more accuracy regarding localization error and also the time required to solve this task is lower than the methods based solely on the nearest neighbour. Among the five holistic descriptors obtained from the CNN, $conv_4$ and $conv_5$ output the best solutions since their localization error as well as their computing time is lower than the obtained through the fully connected layers. These results match the previous conclusion reached in (Cebollada et al., 2019) about the use of 2D convolutional layers to obtain holistic descriptors.

As for the results obtained by conventional and hierarchical localization methods, the conclusion obtained after comparing the results of the table 2 and the fig. 6 is that the hierarchical localization method introduces a faster performance, but it also produces an increase of the localization error. This is due to the fact that the CNN allows a faster rough localization step, but this network produces a small number of wrong room predictions that have a negative influence on the average localization error.

5 CONCLUSIONS

In this work, a study is tackled regarding the use of deep learning to build hierarchical topological models for localization. We also evaluate the ability of the proposed deep learning tool to create holistic descriptors to solve the localization problem based on nearest neighbour. Regarding the hierarchical localization method proposed, this consists in creating a convolutional neural network for classification. This classifier is not only used in the rough localization step to predict the correct room where a test image was captured, but it is also used to obtain a holistic descriptor which characterizes the image. For this work, five layers have been evaluated: $conv_4$, $conv_5$, fc_6 and fc_7 and fc_8. The training and evaluation of all the localization and description methods have been carried out with a panoramic images dataset which contains real conditions effects such as changes in the position of furniture, people walking, blur effect, etc.

As for the use of this CNN to produce global-appearance descriptors for solving the conventional localization by means of a nearest neighbour method, the five descriptors extracted from different layers of the CNN are evaluated together with other analytic holistic methods commonly used for these purposes. The results obtained show that the proposed methods

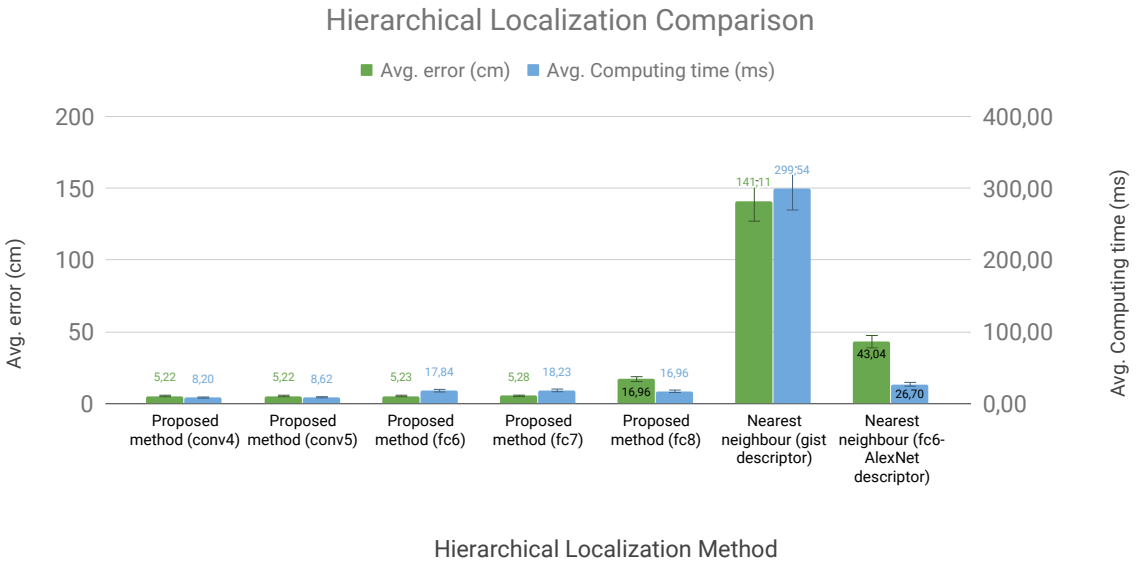


Figure 6: Hierarchical localization methods. Nearest neighbour with either *gist* descriptor or the layer fc_6 of AlexNet and the proposed method based on retrieving the room from the Freiburg CNN and after, solving the fine localization by nearest neighbour with the descriptor obtained from either the layers $conv_4$, $conv_5$, fc_6 and fc_7 or fc_8 of the CNN

are more robust, since they output lower localization error and computing time than the results obtained by analytic methods (*gist* and HOG).

Regarding the hierarchical localization proposed in this work, this has been compared with a method based on obtaining the nearest neighbour through different levels of the model. Prior to this comparison, through the fig. 5, we have showed the accuracy of the trained CNN to estimate the correct room within the environment evaluated. As for the whole localization process, this work shows the evaluation of both methods by using different global-appearance description methods. The method proposed in this paper has proved to be more efficient, since its computing time and localization error are lower than the obtained by means of the nearest neighbour method.

Among the five holistic descriptors obtained from the trained CNN, the descriptor from the layer fc_8 can be discarded, because this descriptor does not characterize properly enough the images for the proposed tasks. The descriptors related to the $conv_4$ and $conv_5$ layers have produced the optimal localization solutions among all the methods evaluated, since the size of the descriptor is relatively small and it leads to low computing time. Despite their size, their localization results are also the most accurate. They produce an average error around 5 cm departing from a training dataset whose average distance between adjacent images is around 20 cm.

In future works, we will spread the evaluation in order to evaluate the goodness of the proposed methods under changes of illumination. Furthermore, we will check whether this CNN is useful to obtain global-appearance descriptors in similar environments. We will also consider other newer and more complex CNN architectures such as ResNet or VGG Net. Last, we would also like to create and evaluate a CNN based directly on omnidirectional images instead of panoramic.

ACKNOWLEDGEMENTS

This work has been supported by the Generalitat Valenciana and the FSE through the grants ACIF/2017/146 and ACIF/2018/224, 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.” and by Generalitat Valenciana through the project AICO/2019/031: “Creación de modelos jerárquicos y localización robusta de robots móviles en entornos sociales”.
The authors declare that there are no competing interests regarding the publication of this paper.

REFERENCES

Abadi, M. H. B., Oskoei, M. A., and Fakharian, A. (2015). Mobile robot navigation using sonar vision algorithm applied to omnidirectional vision. In *2015 AI & Robotics (IRANOPEN)*, pages 1–6. IEEE.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

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.

Cebollada, S., Payá, L., Román, V., and Reinoso, O. (2019). Hierarchical localization in topological models under varying illumination using holistic visual descriptors. *IEEE Access*, 7:49580–49595.

Cebollada, S., Payá, L., Valiente, D., Jiang, X., and Reinoso, O. (2019). An evaluation between global appearance descriptors based on analytic methods and deep learning techniques for localization in autonomous mobile robots. In *ICINCO 2019, 16th International Conference on Informatics in Control, Automation and Robotics (Prague, Czech Republic, 29-31 July, 2019)*, pages 284–291. Ed. INSTICC.

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

Do, H. N., Choi, J., Young Lim, C., and Maiti, T. (2018). Appearance-based localization of mobile robots using group lasso regression. *Journal of Dynamic Systems, Measurement, and Control*, 140(9).

Dymczyk, M., Gilitschenski, I., Nieto, J., Lynen, S., Zeisl, B., and Siegwart, R. (2018). Landmarkboost: Efficient visualcontext classifiers for robust localization. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 677–684.

Guo, J. and Gould, S. (2015). Deep cnn ensemble with data augmentation for object detection. *arXiv preprint arXiv:1506.07224*.

Han, D., Liu, Q., and Fan, W. (2018). A new image classification method using cnn transfer learning and web data augmentation. *Expert Systems with Applications*, 95:43–56.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Lenz, I., Lee, H., and Saxena, A. (2015). Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724.

Liu, R., Zhang, J., Yin, K., Pan, Z., Lin, R., and Chen, S. (2018). Absolute orientation and localization estimation from an omnidirectional image. In *Pacific Rim International Conference on Artificial Intelligence*, pages 309–316. Springer.

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.

Meattini, R., Benatti, S., Scarcia, U., De Gregorio, D., Benini, L., and Melchiorri, C. (2018). An semg-based human–robot interface for robotic hands using machine learning and synergies. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 8(7):1149–1158.

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

Pak, M. and Kim, S. (2017). A review of deep learning in image recognition. In *2017 4th international conference on computer applications and information processing technology (CAIPT)*, pages 1–3. IEEE.

Payá, L., Gil, A., and Reinoso, O. (2017). A state-of-the-art review on mapping and localization of mobile robots using omnidirectional vision sensors. *Journal of Sensors*, 2017.

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.

Pronobis, A. and Jensfelt, P. (2011). Hierarchical multimodal place categorization. In *ECMR*, pages 159–164.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Ullah, M. M., Pronobis, A., Caputo, B., Luo, J., and Jensfelt, P. (2007). The cold database. Technical report, Idiap.

Wozniak, P., Afrisal, H., Esparza, R. G., and Kwolek, B. (2018). Scene recognition for indoor localization of mobile robots using deep cnn. In *International Conference on Computer Vision and Graphics*, pages 137–147. Springer.

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.

Zhu, Y., Mottaghi, R., Kolve, E., Lim, J. J., Gupta, A., Fei-Fei, L., and Farhadi, A. (2017). Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3357–3364.