IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 22, 2019, accepted April 4, 2019, date of publication April 11, 2019, date of current version April 24, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910581

Hierarchical Localization in Topological Models Under Varying Illumination Using Holistic Visual Descriptors

SERGIO CEBOLLADA, LUIS PAYÁ, VICENTE ROMÁN, AND OSCAR REINOSO, (Senior Member, IEEE)
Department of Systems Engineering and Automation, Miguel Hernández University, 03202 Elche, Spain

Corresponding author: Sergio Cebollada (sergio.cebollada@umh.es)

This work was supported in part by the Generalitat Valenciana under Grant ACIF/2017/146 and Grant ACIF/2018/224, and in part by the Spanish government through the project (AEI/FEDER, UE): "Creación de mapas mediante métodos de apariencia visual para la navegación de robots" under Grant DPI 2016-78361-R.

ABSTRACT In this paper, a hierarchical localization framework within indoor environments is proposed and evaluated, considering severe variations of the illumination conditions. The only source of information both to build a model of the environment and to solve the localization problem is a catadioptric vision system, which is mounted on the mobile robot. The images captured by this system are processed globally to obtain holistic descriptors. The position of the robot is estimated by comparing these descriptors with the information contained in a topological visual model, which is previously created using a clustering approach and is composed of a hierarchy of layers. Compacting the information via clustering proves to be an efficient alternative to estimate the position of the robot hierarchically and with robustness. The proposed localization strategy is tested with some sets of panoramic images, captured in large indoor environments under real operating conditions, including illumination changes that change substantially the appearance of the scenes. The results show a reasonable tradeoff computation time-accuracy when the localization is addressed in a hierarchical way.

INDEX TERMS Localization, omnidirectional visual information, global appearance descriptors, clustering, illumination changes.

I. INTRODUCTION

Nowadays, the use of omnidirectional vision sensors in mobile robotics for solving mapping and localization has considerably increased. They have been successfully used by different authors for these purposes. For instance, Valiente *et al.* [1] used the local features extracted from omnidirectional images to generate a reliable visual odometry to improve the Simultaneous Localization And Mapping (SLAM) task. Marinho *et al.* [2] used feature extractions and machine learning techniques to solve localization using omnidirectional images. Faessler *et al.* [3] present a vision-based quadrotor system to map a dense three-dimensional area online with the purpose of removing delay between the quadrotor and external systems. Berenguer *et al.* [4] considered the global appearance of omnidirectional images to create local maps and to estimate the position of a robot within these maps. This kind of images covers a field of view of 360 *deg* around the robot. Hence, they offer a huge amount of information from the surroundings of the robot which permits both building rich maps and estimating the robot position. Working with images requires a step to obtain functional, robust and relevant information from them. Commonly, two methods to extract relevant information have been considered in the related literature: either detecting, describing and tracking some relevant landmarks over the image (such as [5]–[7]) or creating a unique descriptor per image which contains global information about it (for instance, [8]–[10]). As for the second proposed method, on the one hand, it usually leads to more direct localization algorithms. Basically, they consist in a pairwise comparison between descriptors. On the other hand, it presents a lack of metric information. Therefore, this kind of descriptors are usually used to build topological maps (such as [11]–[13]).

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

In order to address the mapping and localization issue, arranging the topological information hierarchically constitutes an efficient alternative. This framework consists in creating a map which is composed of several layers with a hierarchical structure. The high-level ones present a relatively compact amount of information, which permits a rough but quick localization. The low-level layers have usually more information and are used to refine the position. A good example of this issue was developed by Stimec *et al.* [14], who proposed an unsupervised hierarchical mapping method. Garcia-Fidalgo and Ortiz [15] presented a review about the main approaches considered to carry out topological mapping and localization through visual information in the last years. recently, da Silva *et al.* [16] propose a localization and navigation approach for mobile robots using topological maps and using CNN to obtain descriptors from omnidirectional images.

Considering this information, the main objective of this work consists in proposing an approach to solve the localization problem using hierarchical models. Moreover, a comparative evaluation of some global descriptors is carried out to know which one behaves more robustly against illumination changes. The results obtained throughout this work permit selecting the best global descriptor method and also tuning correctly its parameters in order to obtain optimal results (the maximum accuracy and the lowest computational time). Additionally, the use of approaches based on deep learning are also considered to describe the scenes globally. The aim consists in evaluating which method solves more efficiently the localization task under the conditions previously exposed.

An omnidirectional vision sensor [17] is the unique source of information used to carry out mapping and localization in this work. The images used in the experiments are obtained from an indoor dataset (explained in IV-A.1) and they are described through global appearance descriptors. The present work continues and expands the research framework presented in [18], where an approach is proposed to build compact topological models of the environment. The approach consists in the use of clustering algorithms, which are non-supervised techniques, along with holistic visual descriptors, and both the correctness of the model and its utility to solve the localization problem is assessed. An exhaustive evaluation of different clustering methods was carried out in [18]. In that work, Spectral Clustering along with the holistic descriptor *gist* was chosen as the configuration which best tackled the mapping task. Hence, in this work, the localization algorithms are tested with the compact maps obtained with this combination of methods (*gist* + spectral clustering). These compact models are the basis of the present work, whose main differences and contributions are: (a) solving the localization problem hierarchically, with different degrees of granularity, (b) making an exhaustive comparative evaluation of the method and testing its robustness under severe illumination variations and (c) including in the evaluation a new holistic description method, based on deep learning (obtained through convolutional neural networks).

The remainder of the paper is structured as follows. Section II outlines the global appearance descriptors used along this work. After that, section III explains briefly the clustering approach used to compress the information and section IV presents the experiments carried out to test the validity of the proposed methods to solve the localization under changing lighting conditions. At last, the conclusions are presented in section V.

II. THE GLOBAL APPEARANCE DESCRIPTOR

This section focuses on the methods used to describe the global appearance of the set of images. Four methods are evaluated in this paper: the Fourier Signature (FS), the Histogram of Oriented Gradients (HOG), the *gist* of the scenes and a global descriptor based on a Convolutional Neural Network (CNN). In order to reduce the effect of changing lighting conditions, the homomorphic filter [19] is applied over the images before describing them with HOG, since previous works [20] concluded that this pre-filtering improves the localization results when HOG is used.

The panoramic image $im(x, y) \in R^{N_x \times N_y}$ is the starting point, hence, a conversion from omnidirectional to panoramic must be carried out. After that, one of the four proposed description methods is used to calculate the global appearance descriptor vector $d \in R^{l \times 1}$. A deep description of FS, HOG and *gist* methods can be found in [21]. As for the use of CNN as global feature extractor, a wide explanation is presented in [22].

Regarding the FS descriptor, it was firstly used by Menegatti *et al.* [8]. This method calculates the discrete Fourier Transform of each row of the panoramic image and a complex matrix is obtained $IM(u, v)$. The k_1 first columns are retained (compression effect) $IM(u, v) \in C^{N_x \times k_1}$. Finally, a decomposition is tackled to obtain just the magnitudes information (the resulting matrix is invariant to robot orientation changes) and the rows of the resultant matrix are arranged to create a vector, obtaining the global appearance descriptor $d \in R^{N_x \cdot k_1 \times 1}$.

As for the HOG descriptor, it was firstly used by Dalal and Triggs [23] for a pedestrians detection task. The version used in this work consists in splitting the panoramic image into k_2 horizontal cells and compiling a histogram of gradients orientation per each cell with b bins per histogram [24]. The set of histograms compose the final descriptor $d \in R^{b \cdot k_2 \times 1}$.

With regard to the *gist* descriptor, Oliva and Torralba [25] introduced this method, which has been widely used for scenes recognition. Several versions can be found depending on the features of the image used. In this case, firstly, m_2 different resolution images are created from the original panoramic one. Secondly, Gabor filters are applied over the m_2 images with m_1 different orientations each. Thirdly, the pixels of each image are grouped into k_3 horizontal blocks and finally, the obtained orientation information is grouped to create a vector, which is the resultant descriptor $d \in R^{m_1 \cdot m_2 \cdot k_3 \times 1}$. This descriptor has already been used in mobile robot localization. For instance, Murillo *et al.* [26] used

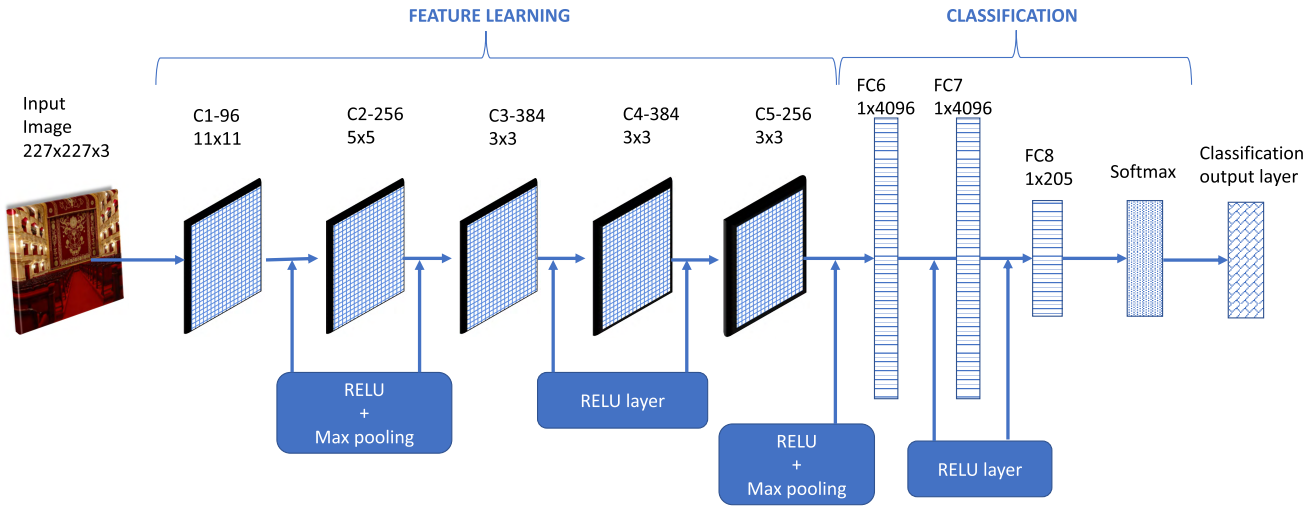


FIGURE 1. CNN *places* architecture design based on the pre-trained *'Caffe'* model. Layers *'fc7'* and *'fc8'* are used in this work as a method to obtain holistic descriptors from the original input image.

it with panoramic images for localization in urban regions including loop closure detection.

Last, concerning the use of the CNN-based descriptor, this method comes from the use of deep learning for classification, as Krizhevsky *et al.* [27] do. The neural network tackles two steps. First, it carries out a learning process, i.e., a set of images (which are already labeled) are collected and introduced to the network. Second, once trained, the network receives new images (also labeled) and tunes its internal parameters to optimize the results. After that, the network is available to face the classification task: a new image is introduced and the CNN returns the most likely label option. During the process of classification, descriptors are obtained by the fully connected layers which are within the neural network. These descriptors can be seen as global appearance descriptors of the input image. Therefore, they may be also used to carry out the localization task in the same way as the previously proposed global appearance descriptors. The neural network architecture that we use in this work is *places* [28], which was trained with around 2.5 million images to categorize 205 possible kinds of scenes. The fig. 1 shows the architecture of this CNN. To obtain holistic descriptors from these layers, the networks is directly used with the pre-training done by the creators, hence, a re-training is not necessary. The CNN is used directly as it appears in [29]. The descriptors extracted from this network correspond to the ones calculated in the layers *'fc7'* and *'fc8'*. These descriptors contain respectively 4096 ($d \in R^{4096\times 1}$) and 205 ($d \in R^{205\times 1}$) components. This kind of descriptor has been used by other authors such as Mancini *et al.* [30], who use them to carry out place categorization with the Naïve Bayes classifier. As for mobile robot localization, Payá *et al.* [22] proposed CNN-based descriptors to create hierarchical visual models. In a different way, Xu *et al.* [31] propose the use of a CNN which detects objects from the images and establishes relationships between the detected objects. Afterwards, the relationships established are used to calculate similitude between images. Nevertheless, in our work, CNNs are used just with the purpose of obtaining a holistic descriptor per scene.

III. CLUSTERING THE VISUAL INFORMATION

This section outlines the clustering method used to compact the model. The clustering process departs from a set of images $I = \{im_1, ..., im_n\}$. These images were captured from different positions within the environment to map. The image capturing positions are known, but they are only used as *ground truth* $P = \{(x_1, y_1), ..., (x_n, y_n)\}$. Then, a set of global appearance descriptors $D = \{d_1, ..., d_n\}$ is calculated, one per image (through one of the description methods explained in section II). To create a compact model, a clustering process will be carried out with the components of D.

Several studies about clustering have been carried out. For instance, Theodoridis and Koutroumbas [32] developed a wide study about clustering and von Luxburg [33] provided a complete tutorial about the most common spectral clustering methods. This kind of algorithms have proved to be more effective than the traditional ones when the data size is high. Furthermore, the Spectral Clustering developed by Ng *et al.* [34] confirmed to be a good solution in these situations. This algorithm only considers the similitudes between instances d_i and d_j: $S_{i,j} = e^{-\frac{|d_i - d_j|^2}{2\sigma^2}}$, where σ is a parameter which controls the rapidity of reduction of the similitude when the distance between d_i and d_j increases. The clustering process is as follows:

1) Calculation of the normalized Laplacian matrix:

$$L = I - D^{-1/2} S D^{1/2} \tag{1}$$

where D is a diagonal matrix $D_i = \sum_{j=1}^{N} S_{ij}$.

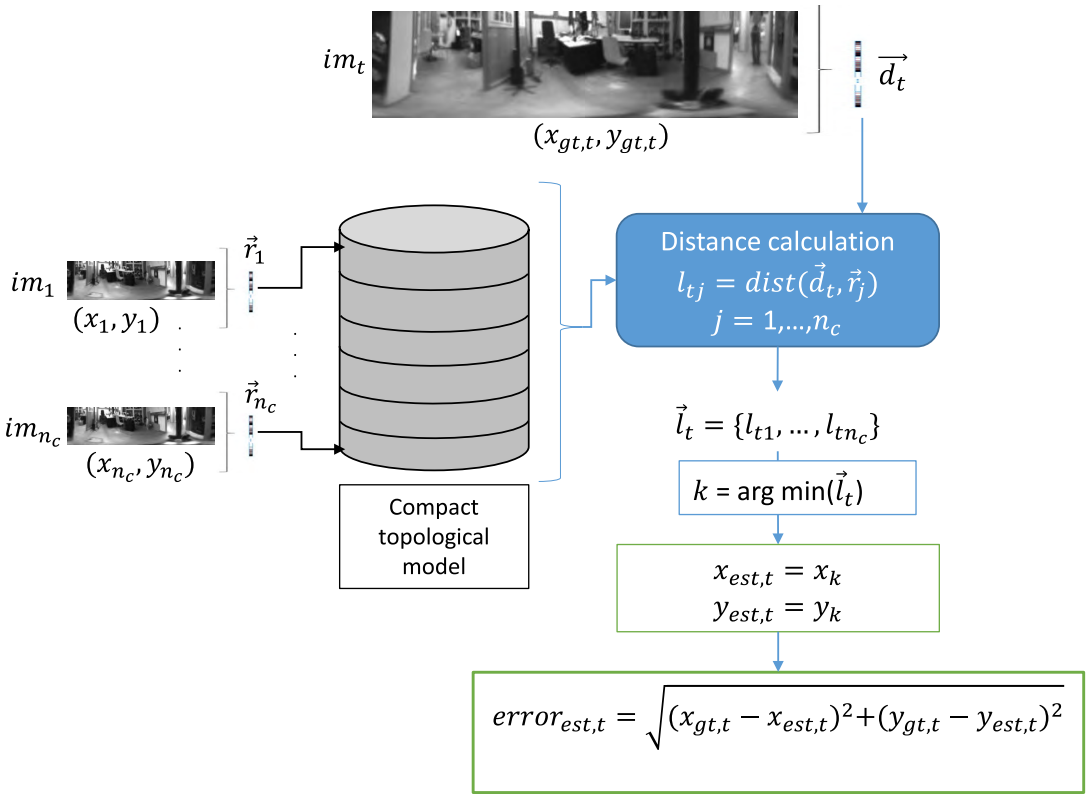


FIGURE 2. Block diagram regarding the steps to carry out the localization task through compact models.

2) Calculation of the n_c main eigenvectors of L, $\{u_1, u_2, ..., u_{n_c}\}$. Arranging these vectors by columns, the matrix $U \in R^{n \times n_c}$ is obtained.
3) The matrix U is normalized to obtain the matrix $T \in \mathbb{R}^{n \times n_c}$.
4) Extraction of vector $y_i \in R^{n_c}$ from the i-th row of the matrix T. $i = 1, ..., n$.
5) The y_i vectors are clustered by using a simple clustering algorithm (k-means in this work). The clusters $A_1, A_2, ..., A_{n_c}$ are obtained.
6) Last, the clusters with the original data are obtained as $C_1, C_2, ..., C_{n_c}$ where $C_i = d_j$ such that $y_j \in A_i$.

After the clustering process, the representative of each cluster is calculated as the average of the descriptors which compose a specific cluster. The final result is a set of representatives $R = \{r_1, ..., r_{n_c}\}$, which constitutes the compressed map (i.e. the high-level layer of the hierarchical map). It can be used to carry out the localization task in a more efficient way.

IV. LOCALIZATION UNDER CHANGING LIGHTING CONDITIONS

In a previous work [20], among the traditional global appearance descriptors, *gist* proved to be the most efficient to compact the model in indoor environments. Now, a comparative study of the proposed descriptors for localization under illumination changes is tackled including also the description method based on CNN.

A. LOCALIZATION THROUGH COMPACT MODELS

The localization step is carried out once the compressed map is built. Hence, the starting point is a compact topological model, which consists of a set of n_c representatives $\{r_1, ..., r_{n_c}\}$ and the coordinates of each cluster $\{(x, y)_1, ..., (x, y)_{n_c}\}$, where n_c is the number of clusters. Nevertheless, the coordinates are only used as ground truth to test the accuracy. Only visual information is used during the localization. It allows us to carry out a pure evaluation of the visual description methods through avoiding the influence of other type of information. The accuracy is evaluated through the following error equation $error_{est,t} = \sqrt{(x_{gt,t} - x_{est,t})^2 + (y_{gt,t} - y_{est,t})^2}$, where $(x_{gt,t}, y_{gt,t})$ is the pose provided by the ground truth and $(x_{est,t}, y_{est,t})$ is the pose estimated by the algorithm for the test image t.

The localization task is performed through the following steps: first, it is assumed that the previous position of the robot is unknown; second, the robot captures a new image im_t (an image from the test dataset which is different from the images used to create the map) and describes that image to obtain the descriptor d_t; third, the distance between d_t and each representative descriptor is calculated, obtaining a distances vector $l_t = \{l_{t1}, ..., l_{tn_c}\}$ where $l_{tj} = dist(d_t, r_j)$; fourth, the minimum value of l_t indicates the cluster which corresponds to the current position of the robot. A block diagram about these steps is shown in fig. 2.

1) EXPERIMENTS

To carry out the experiments, the COLD (COsy Localization Database) database is used [35]. This database is composed of several sets of omnidirectional images captured with a catadioptric vision system composed of a *Videre Design MDCS2* camera and a hyperbolic mirror. The images were collected under three different illumination conditions (cloudy days, sunny days and at nights). The Freiburg and Saarbücken datasets (images acquired at indoor laboratory environments located in those cities) were used to develop the experiments in this work. The images captured during cloudy weather are used to build a compact model through spectral clustering since they are the ones which are less affected by brightness, reflections, dark areas and thus, they provide more information. The sunny weather images and also the images captured at night are used as test images to evaluate the localization task under lighting changes.

Both datasets are composed of several rooms, such as corridors, personal offices, printer areas, kitchens, bathrooms, etc. The selected Freiburg dataset covers 9 different rooms and the Saarbrücken dataset covers 8. This dataset includes different changes in the environment such as people walking or position of furniture and objects. The datasets contain also images which do not provide much information due to the acquisition position and blurry images. All these handicaps make these datasets suitable to carry out experiments under real operating conditions. From the original cloudy dataset, a downsampling is carried out in order to obtain an acquisition distance between images of 40 cm approximately. This downsamplig is carried out because it is desirable to keep the model configuration which was used in previous works ([20] and [21]). Hence, after downsampling, a training dataset composed of 519 (in Freiburg) and 566 (in Saarbrücken) images are considered. Furthermore, departing from the sunny and night datasets, three test datasets are created. Those images were selected randomly across the whole map. The table 1 summarizes the sets created for the experiments. The fig. 3 shows some examples of omnidirectional images in both environments under the proposed illumination conditions.

TABLE 1. Datasets created from the COLD database to carry out the experiments.

Dataset name	Illumination condition	Number of images	Path length (m)
Freiburg_training	Cloudy	519	104.2
Freiburg_test_night	Night	58	
Freiburg_test_sunny	Sunny	45	
Saarbrücken_training	Cloudy	566	156.6
Saarbrücken_test	Night	57	

As mentioned before, the localization experiment departs from the compact model. Several compact maps have been built, considering different numbers of clusters $n_c = [10, 20, ..., 100, N_{env}]$ where N_{env} is the total number of images which compose the model (519 in the Freiburg environment and 566 in Saarbücken, table 1). It will enable us to analyze the localization process considering different granularities in the high-level layer of the map. The case $n_c = N_{env}$ provides information about the localization process when all the images of the original model are considered (i.e., no compression is performed and the localization is addressed as an image retrieval problem). This way, it can be seen as a reference to test the utility of the compact maps. To create the clusters, *gist* was chosen since it has provided the best results in previous works [20] and its parameters are tuned to $k_3 = 32$ and $n_{masks} = 16$. The fig. 4 shows examples of a sample clustering experiment applied to the datasets, according to the spectral clustering method. The images of these datasets are under cloudy illumination conditions.

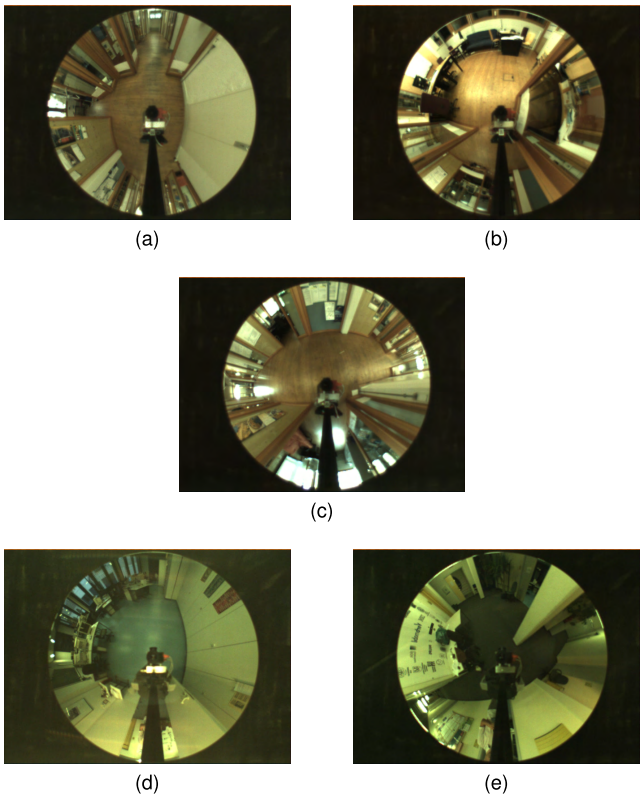


FIGURE 3. Some sample omnidirectional images belonging to the Freiburg environment under (a) cloudy, (b) night and (c) sunny illumination conditions and also images which belong to the Saarbrücken environment under (d) cloudy and (e) night illumination conditions.

Once the compact map is available (i.e. the clusters' representatives have been calculated), the localization is estimated as follows. Among the n_c clusters, the node whose distance presents the minimum value of l_t is chosen as the one which the captured image belongs to. Therefore, to estimate the goodness of the localization task, the Euclidean distance between the position where the test image was captured and the position of the nearest neighbour is calculated. Additionally, the computation time is measured since the scope is to reach a balance between accuracy and computational time. The experiments have been carried out in a PC with two CPU Quad-Core Intel Xeon® at 2,8 GHz and through Matlab® programming.

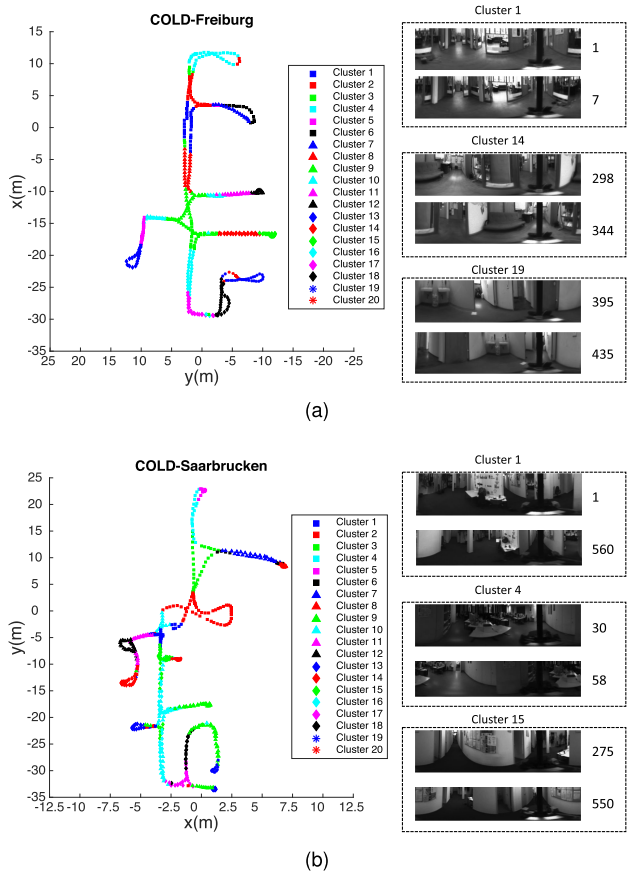


FIGURE 4. Results of a sample clustering considering $n_c = 20$ clusters and *gist* descriptor with $k_3 = 32$ and $n_{masks} = 16$. Some sample panoramic images belonging to the (a) Freibug and (b) Saarbrücken dataset under cloudy illumination conditions are shown.

To calculate the distance between descriptors, three kinds of distances are considered: the correlation distance, the cosine distance and the Euclidean distance. Also, three illumination conditions are considered: the cloudy condition, the night condition and the sunny condition. The dataset under cloudy condition is the one used to create the compact map (clustering and obtaining the representatives). Night and sunny conditions are used to evaluate the localization task under illumination changes. Fig. 5 shows the average localization error (cm) vs. the number of clusters n_c obtained in the Freiburg environment when the test dataset was night and fig. 6 when the test dataset was sunny; fig. 7 shows the average localization error (cm) obtained in the Saarbrücken environment when the test dataset was night. In all cases, the localization error is expressed in cm, and the colorbar that expresses this error has the same range, to facilitate a comparative evaluation between figures.

In general, as the number of clusters increases, the average localization error tends to decrease. This behavior was expected and was also remarked in previous works [20]. When there is a low number of clusters, the plots present high error values, as expected. This is due to the fact that

TABLE 2. Computation time (sec) required to obtain the global appearance descriptor (HOG and CNN) per each test image. Freiburg test dataset under night conditions.

descriptor		time (ms)
HOG	k2=2	131.0 ± 0.58
	k2=4	145.8 ± 0.34
	k2=8	148.3 ± 0.12
	k2=16	158.1 ± 0.19
	k2=32	177.8 ± 0.93
	k2=64	192.3 ± 0.41
CNN	'fc7'	444.7 ± 5.62
	'fc8'	453.3 ± 4.51

despite the matching between test images and representatives has been successful, the representatives are too sparse among them. Moreover, as for the illumination conditions, if we compare the outputs obtained under night conditions and the ones obtained under sunny conditions (see fig. 5 and fig. 6), generally, sunny conditions have a more negative impact upon the localization. For example, when using the CNN descriptor layer *'fc7'*, if $n_c = 10$, the error under night conditions is over 200 cm whereas under sunny conditions, it is over 300 cm. If $n_c = 60$, the error under night conditions is under 100 cm and under sunny conditions, it is over 200 cm.

Among the four studied global appearance descriptors, FS is the one which presents worst localization results in general. As for HOG, this descriptor presents relatively good localization error results. For example, for night conditions in the Freiburg environment (see fig. 5), when a correct tuning of the k_2 parameter is carried out and for more than 50 clusters, the localization error values are under 100 cm. It can be considered a successful result considering the size of the environment (table 1) and the granularity of the compact map. Moreover, the best results are obtained when the cosine distance is applied. In the case of the Saarbrücken environment, HOG presents slightly worse results than in Freiburg (fig. 7). *Gist* presents also relatively good localization results and they are not as influenced by the k_3 parameter (number of horizontal blocks) as the HOG results with k_2. For instance, in the *gist* results presented in the fig. 5, the error decreases until the number of clusters is 40 and after that value, the average localization error keeps almost constant. For the *gist* descriptor, the Euclidean distance presents the worst results whereas the cosine and correlation distances are quite similar. The CNN descriptor presents as good localization results as using HOG in Freiburg at night. Through the use of CNN with the layer *'fc7'*, the localization error is lower than 100 cm when $n_c > 30$ (using either correlation or cosine distance). Moreover, the results in the Saarbrücken environment are the best. Nevertheless, under sunny conditions (see fig. 6), CNN is more affected than HOG.

Among the two best descriptors which present best localization outputs (HOG and CNN), a computation time evaluation is obtained. With this aim, the time required to calculate the global appearance descriptors is performed. Table 2 shows

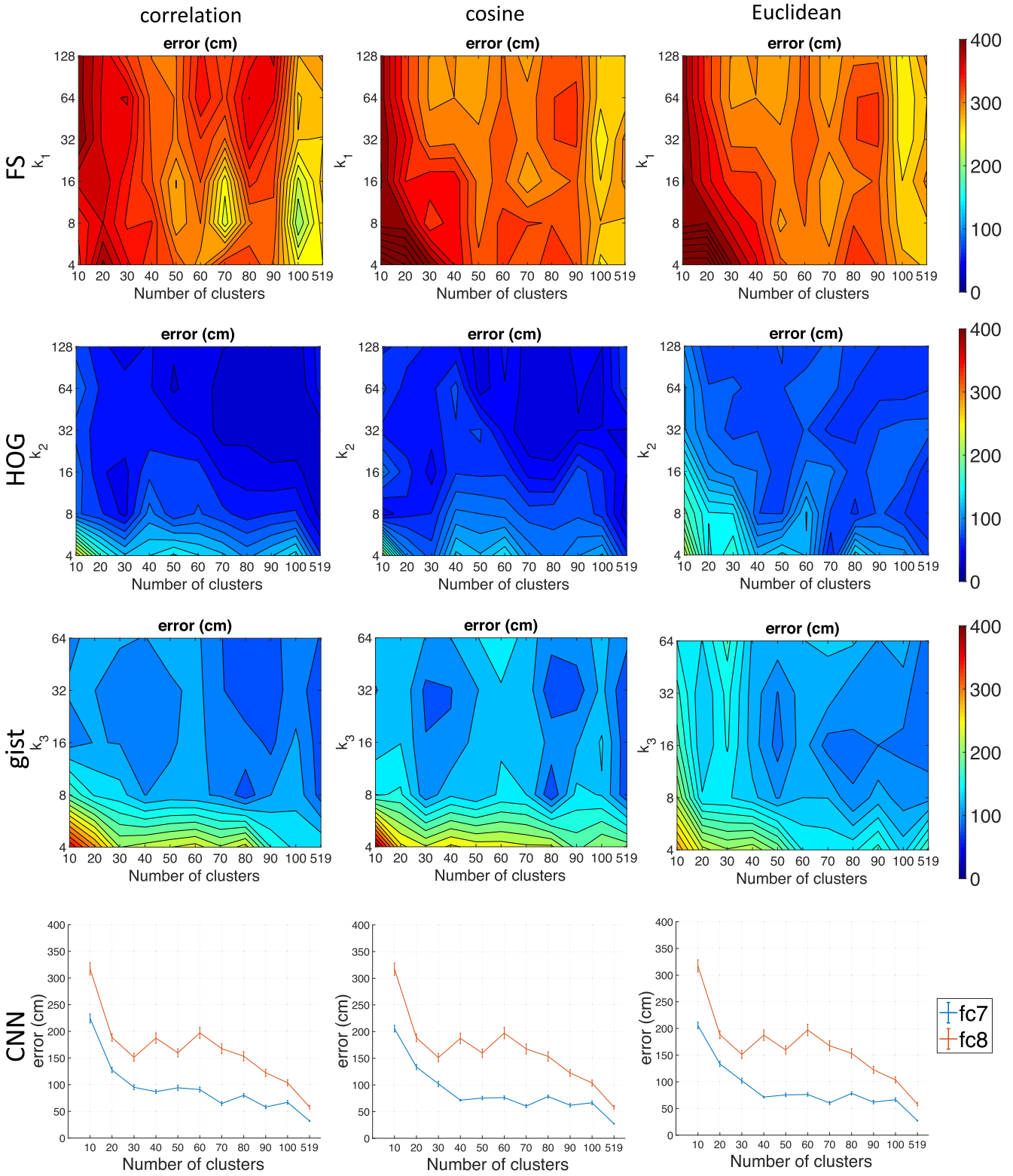


FIGURE 5. Results of the localization task when the night illumination conditions affect the Freiburg environment. Average localization error (cm) vs. number of clusters and descriptor size. Different description methods (FS, HOG, *gist* and CNN) and distances (correlation, cosine and Euclidean) are considered.

the average computational time (sec) to compute the global appearance descriptor for the *Freiburg_test_night* dataset. As for HOG, the obtained values keep almost constant independently on the value of k_2. Regarding the use of the CNN descriptor, the related time values are higher (around 0.4 sec).

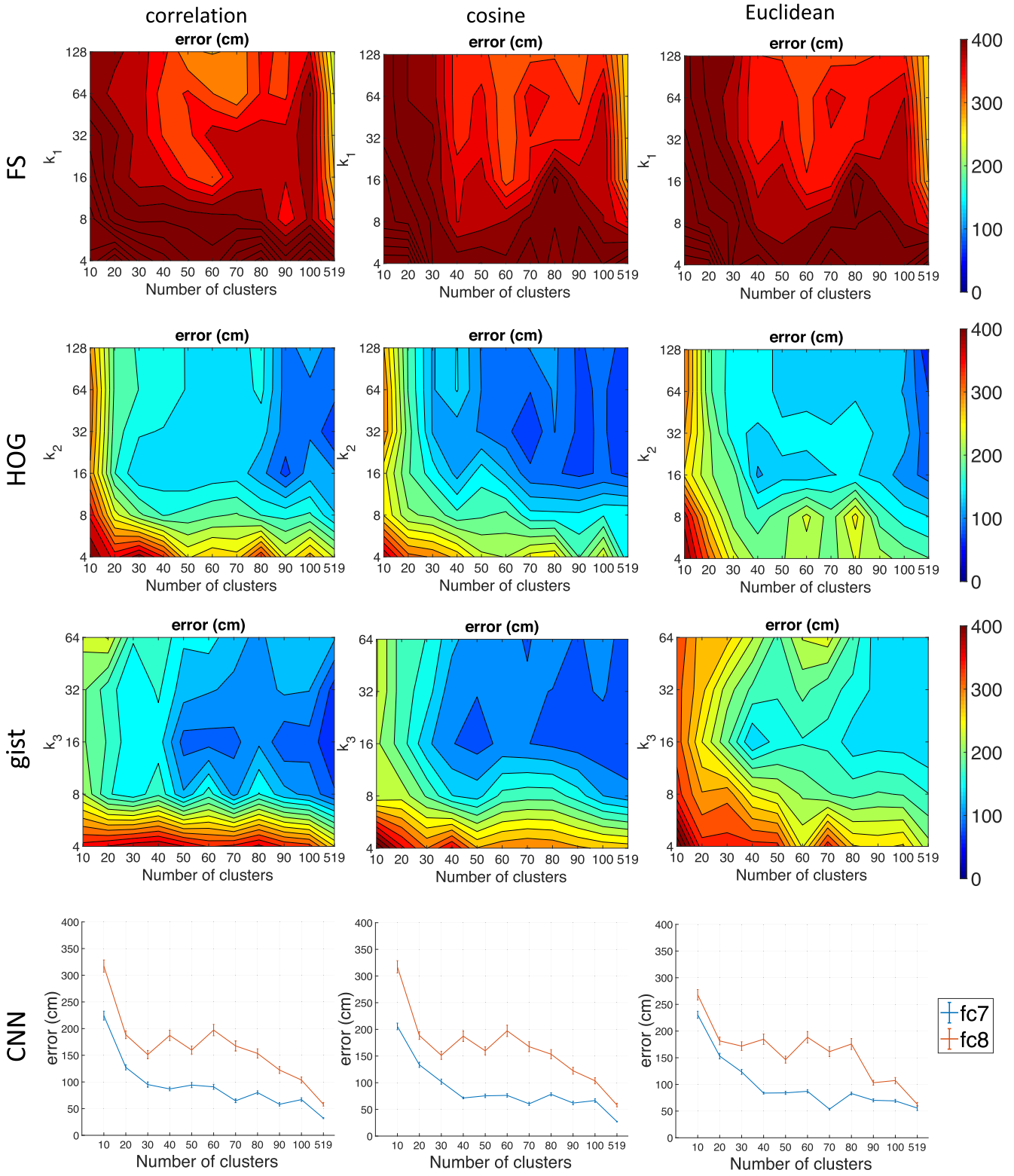


FIGURE 6. Results of the localization task when the sunny illumination conditions affect the Freiburg environment. Average localization error (cm) vs. number of clusters and descriptor size. Different description methods (FS, HOG, *gist* and CNN) and distances (correlation, cosine and Euclidean) are considered.

In conclusion, among the different global appearance descriptors studied to solve the localization task in environments which present changes of illumination, CNN will be the optimal option. FS and *gist* localization values are relatively worse. HOG presents better results in the Freiburg environment, but in the Saarbrücken environment, results for

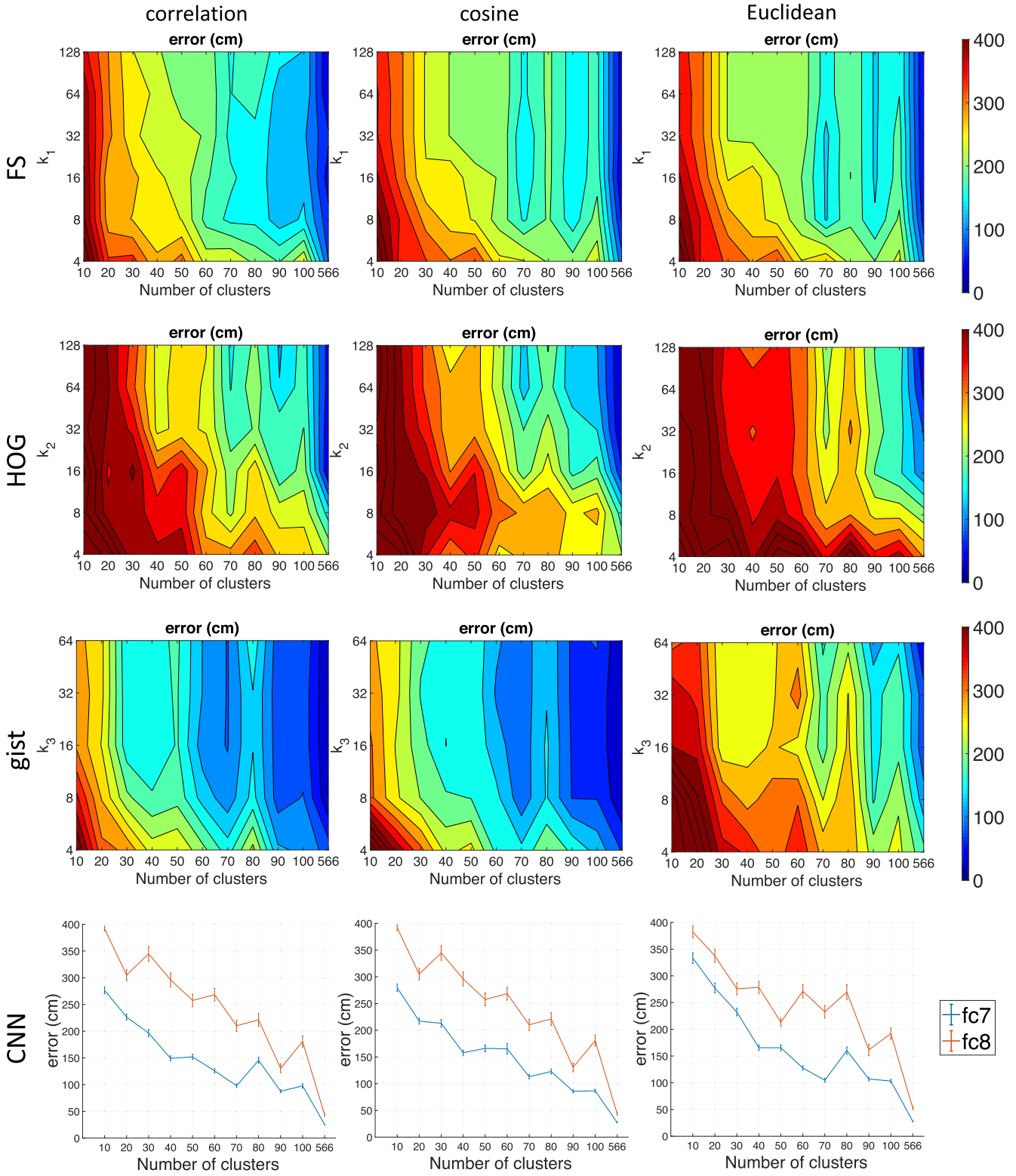


FIGURE 7. Results of the localization task when the night illumination conditions affect the Saarbrücken environment. Average localization error (cm) vs. number of clusters and descriptor size. Different description methods (FS, HOG, *gist* and CNN) and distances (correlation, cosine and Euclidean) are considered.

HOG are poor, whereas CNN keeps being also good. Despite the computing time is not as low as the HOG one, it is not substantially higher than the HOG results. As for the illumination changes, HOG is less affected by the sunny conditions than the rest of descriptors. Regarding which type of distance measure is better to calculate the distance

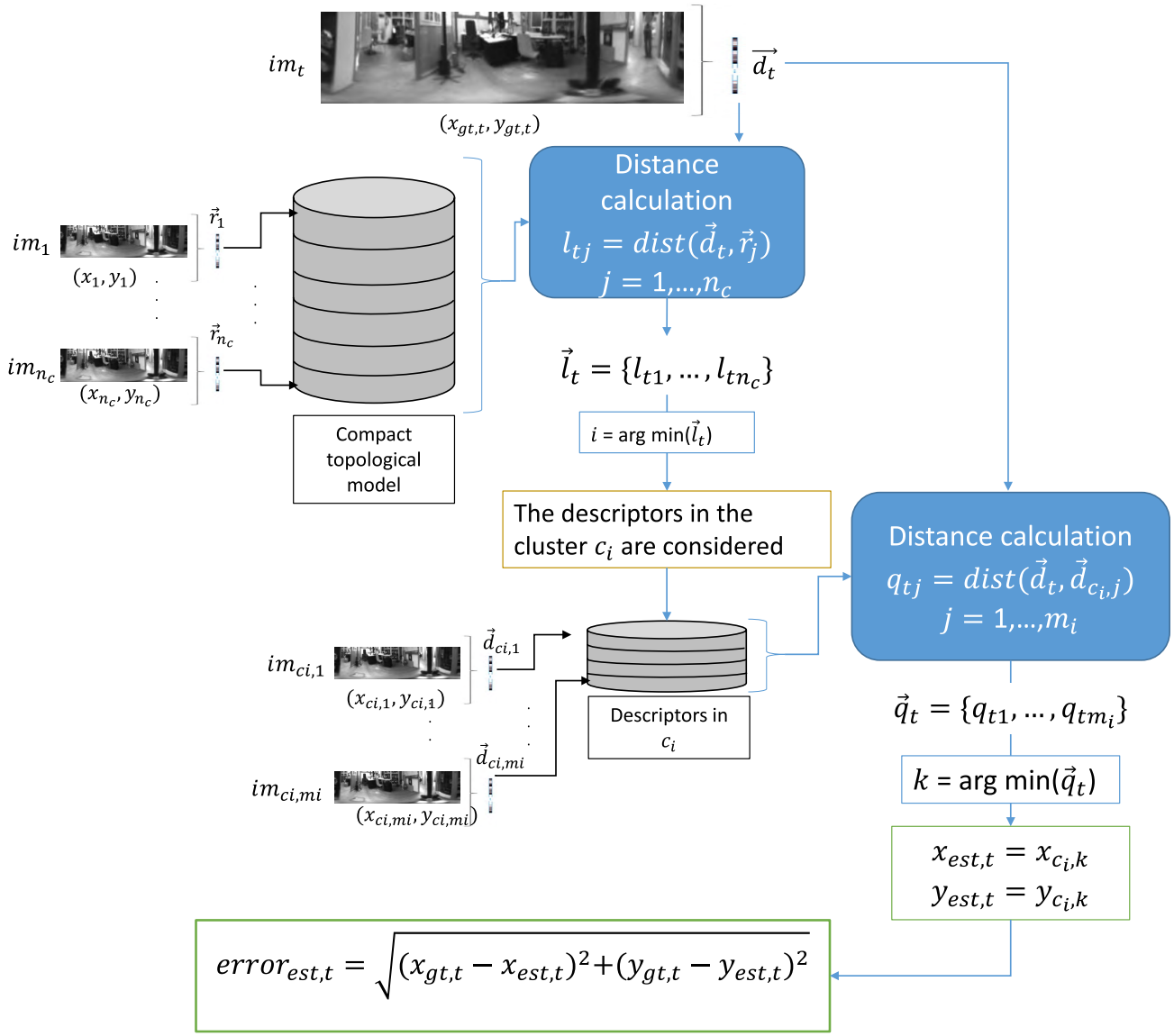


FIGURE 8. Block diagram regarding the steps to carry out the hierarchical localization task through compact models.

between descriptors, both corrrelation and cosine present similar outputs.

B. HIERARCHICAL LOCALIZATION

In subsection IV-A, the localization has been solved using only the compact map (i.e., only the high level layer is used, and the result is a coarse localization). It has allowed us to analyze how different compression levels have an influence on the localization error (i.e. the tradeoff map granularity - localization accuracy).

In this subsection, we go one step beyond, and the localization is addressed hierarchically. First, a coarse localization is performed, as in subsection IV-A. Once the nearest cluster has been retrieved, a second step is carried out to refine the estimation.

Therefore, the hierarchical localization task consists of the following processes: first, the robot describes the image captured at time instant t (test image) $im_t \rightarrow d_t$. After that, the distances vector is again obtained $l_t = \{l_{t1}, ..., l_{tn_c}\}$. Next, the most likely cluster is selected as the one which presents the minimum value of l_t. At this step, a new comparison is carried out between the descriptor of the test image d_t and the descriptors of the images which belong to the chosen cluster. From this step, a new distances vector is obtained $q_t = \{q_{t1}, ..., q_{tm_i}\}$ where m_i is the number of images within the selected cluster i. Finally, the minimum value of q_t indicates the most similar image and hence, it corresponds to the current position of the robot with a higher accuracy. Fig. 8 shows the block diagram about these steps. It should be mentioned that more than one cluster may be selected. The higher

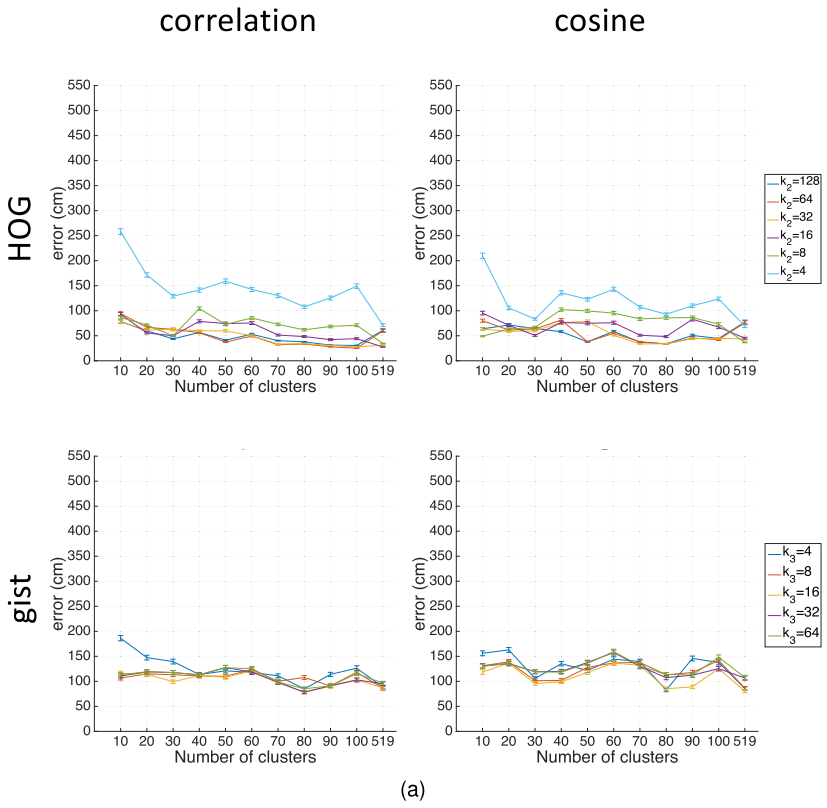


FIGURE 9. Results of the complete hierarchical localization task when the night condition of illumination is affecting the Freiburg environment. Average localization error (cm) vs. number of clusters and descriptor size. Pre-selection of either (a) one ($c = 1$) or (b) two ($c = 2$) clusters as the most likely options.

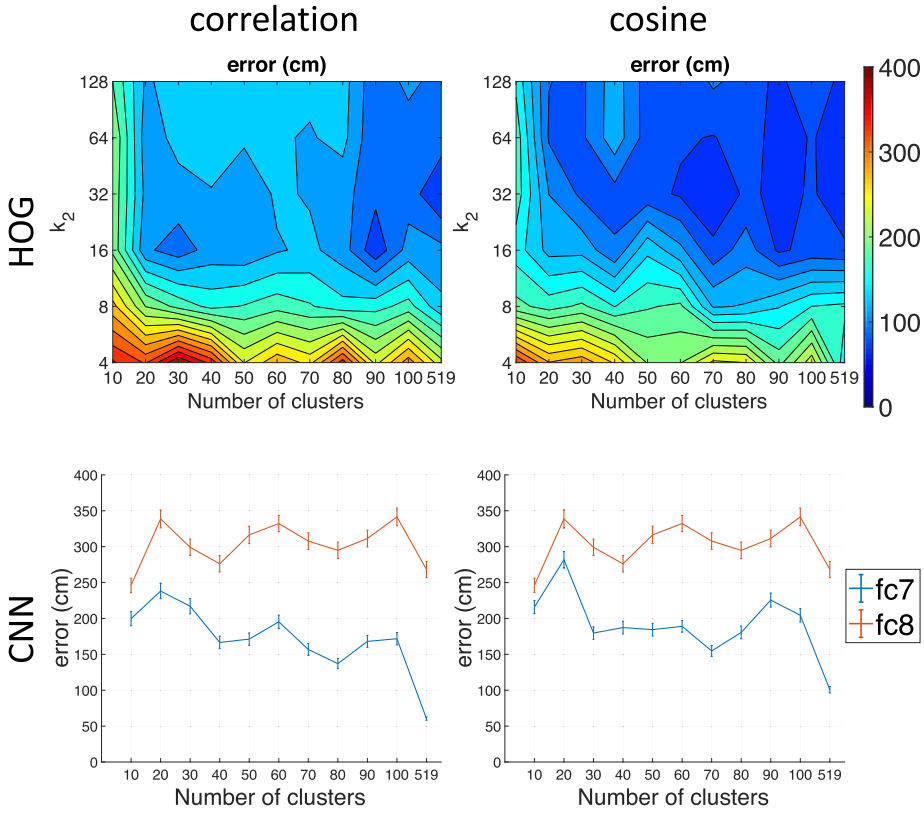


FIGURE 10. Results of the complete hierarchical localization task when the sunny condition of illumination is affecting the Freiburg environment. Average localization error (cm) vs. number of clusters and descriptor size. Pre-selection of one cluster as the most likely option.

the number of selected clusters, the more comparisons with images will be tackled.

1) EXPERIMENTS

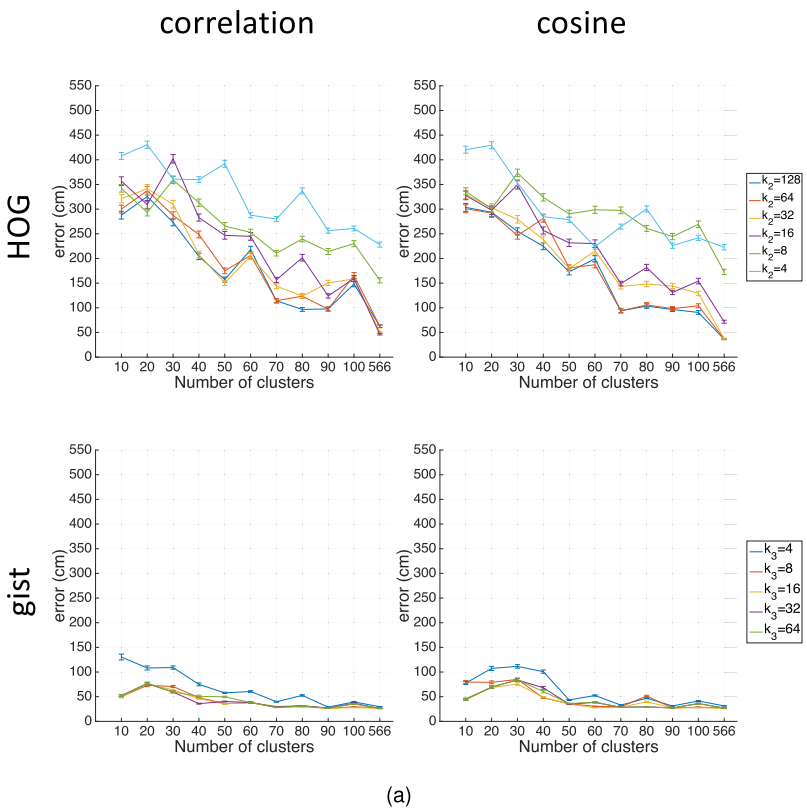
As in the sub-subsection IV-A.1, the experiments were carried out through the use of the COLD database with the same characteristics previously commented. Again, the starting point of the localization experiment is the compact model through *gist* ($k_3 = 32$ and $n_{masks} = 16$).

Since the Euclidean distance presented the worst localization error results in the experiment 1, this distance is discarded. Furthermore, neither FS nor *gist* descriptor related results are shown in this experiment because, as was shown in the previous subsection, those results are worse for localization purposes. Therefore, to sum up, the hierarchical localization is evaluated in the Freiburg and Saarbrücken environments under two illumination conditions (night and sunny) calculating two types of distances (correlation and cosine) and using two kind of descriptors (HOG and CNN).

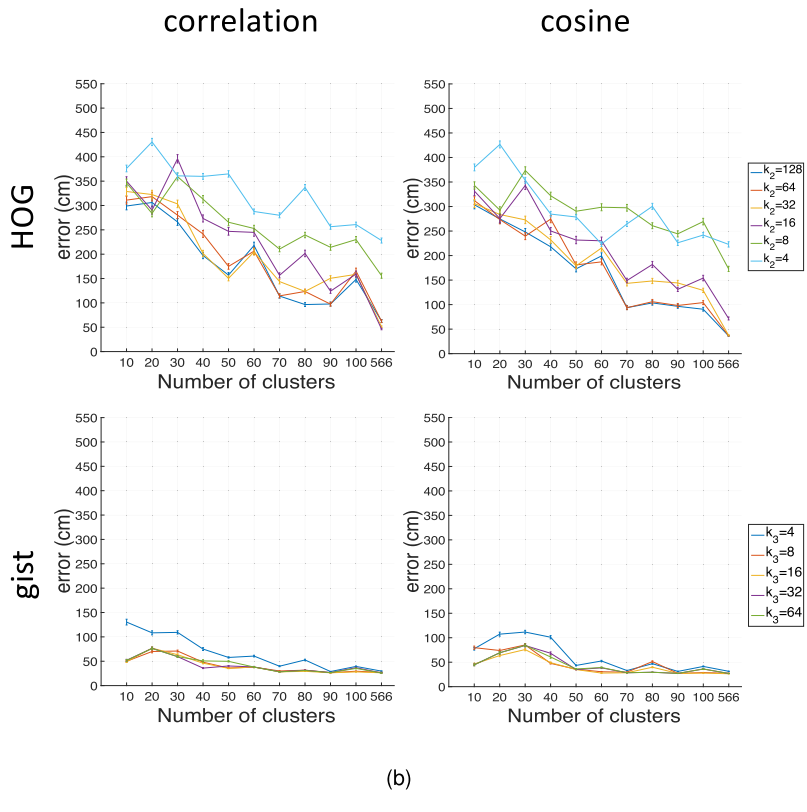
Fig. 9 shows the average localization error (cm) vs. the number of clusters n_c obtained in the Freiburg environment when the test dataset was night and either one or two clusters are selected to carry out the fine localization. Fig. 10 shows the average localization error (cm) obtained in the Freiburg environment results when the test dataset was sunny and one cluster is selected for fine localization; fig. 11 shows the average localization error (cm) vs. the number of clusters n_c obtained in the Saarbrücken environment when the test dataset was night and one and two clusters are selected to carry out the fine localization.

The evaluation of these results is carried out from three points of view. Firstly, a comparison of the results obtained through hierarchical localization against the localization tackled in the subsection IV-A.1 is performed. In general, the localization error obtained through hierarchical localization clearly improves when the number of clusters is low (see fig. 9, 10 and 11). However, when the number of clusters increases, no improvements are noticed. This behaviour is due to the fact that a low number of clusters implies a very rough initial localization. This way, the second step really permits refining this estimation. However if the number of clusters is relatively high, the initial estimation is quite fine and the improvement achieved in the second step is not substantial. This allows us to conclude that the high-level layer can be built with a high degree of compression and the localization can be refined with the low-level layer, through an efficient process.

Secondly, after proving that hierarchical localization presents improvements when a high compression is carried out, an analysis about how illumination conditions affect to



(a)



(b)

FIGURE 11. Results of the complete hierarchical localization task when the night condition of illumination is affecting the Saarbrücken environment. Average localization error(cm) vs. number of clusters and descriptor size. Pre-selection of either (a) one ($c = 1$) or (b) two ($c = 2$) clusters as the most likely options.

TABLE 3. Summary of the minimum localization error values obtained through the two localization methods and the four descriptors evaluated throughout this work.

Localiz. Metthod	Descriptor	Minimum localization error (cm)				
		$n_c = 20$	$n_c = 40$	$n_c = 60$	$n_c = 80$	$n_c = 100$
Localization through compact models	Fs	403.25	331.88	350.88	351.81	267.60
	HOG	149.75	128.19	118.53	112.20	90.85
	gist	201.33	125.77	149.45	125.92	153.38
	CNN	133.54	71.39	76.23	78.27	66.51
Hierarchical Localization	Fs	360.73	308.40	327.13	328.32	252.17
	HOG	61.41	81.27	54.89	33.59	42.52
	gist	136.27	99.10	136.58	85.02	125.31
	CNN	69.52	38.68	50.57	49.01	50.73

hierarchical localization is tackled. Comparing the results obtained when a hierarchical localization process is developed under night conditions (see fig. 9, $c = 1$) and sunny conditions (see fig. 10); several conclusions can be extracted. For a localization task carried out through the use of HOG descriptor and correlation distance in Freiburg under night conditions, the average localization error is between less than 50 and 250 cm, whereas, under sunny conditions, this value is between 70 and 400 cm. This analysis can be extended to the CNN descriptor, which is again highly affected by the sunny conditions. Therefore, collecting results from both experiments, the conclusion is that the sunny condition affects to a greater extent the localization task.

Thirdly, an evaluation about varying the number of clusters selected to carry out the fine localization is done. If we compare the hierarchical localization results in Freiburg for one selected cluster and the results for two selected clusters (see fig. 9), a slight improvement is appreciated in the case $c = 2$ when the number of clusters is low. This behavior means that for few clusters, selecting the right one can be more challenging. Hence, selecting more than one cluster for fine localization may result beneficial when a huge compression was carried out. For the Saarbrücken environment, no improvements have been noticed between selecting one and selecting two clusters (see fig.11). This lack of improvement means that the instances are very well represented even when there is a high level of compression and thus, selecting more than one cluster does not provide a higher probability to find the more accurate position of the test image.

C. DISCUSSION OF RESULTS

The scope of these experiments is to evaluate the robustness of global appearance descriptors to solve the localization problem using hierarchical maps either (1) by using a localization method which estimates the position through compact models, or (2) by solving also a fine localization step (hierarchical localization method). For the sake of completeness, the experimental section considers several methods to obtain the global appearance descriptors (FS, HOG, *gist* and CNN), different configuration parameters of these descriptors and also a variety of illumination conditions. Regarding the localization method through compact models, the average accuracy improves as less compression is tackled. As for the hierarchical localization, this method produces an efficient process to refine the localization in the low-level layer. Nevertheless, this method only improves when the number of clusters is low but no substantial differences exist when the number of clusters is relatively high. Selecting more than one cluster to carry out the fine localization is only interesting when a huge compression is carried out, otherwise, selecting only one cluster produces more efficient results because its computing time is relatively low.

Concerning the global appearance descriptors, FS always outputs the worst results and HOG usually leads to the best solutions. We have found out that the CNN-based descriptor also presents good results. Nevertheless, CNN is more affected by the sunny illumination conditions than HOG is and CNN also needs more computing time to calculate the descriptor than HOG. In general, the sunny illumination conditions affect more negatively the performance of the methods than the night conditions.

V. CONCLUSION

In this work, a study is carried out about the utility to solve the localization task hierarchically in mobile robotics when substantial illumination changes are present. This task is tackled once a compact model of the environment is created. Two indoor sets of 519 and 566 panoramic images have been respectively used. A clustering approach through Spectral Clustering with a number of clusters between 10 and the total number of instances was considered. Therefore, a reduction between 1.77% and 17.67% of information contained in the initial set of images is considered. Additionally, we also analyze the localization when no compression is done, as a reference.

The work has shown that it is possible to keep a good localization error departing from a compact model. The issue is solved through the use of global appearance of panoramic scenes. A comparative evaluation between four methods to globally describe images has been carried out: FS, HOG, *gist* and a CNN-based descriptor. The CNN-based descriptor and cosine distance has been proved to be the best choice. The table 3 summarizes the localization error obtained along this work. Through this table, it is easy to conclude that the CNN-based descriptor provides the best results to carry out

the localization task for both localization methods although HOG also presents good results when the localization is addressed hierarchically.

This work has also shown the efficiency of this localization framework under severe changes of illumination. Moreover, it has proved that the test images under sunny conditions affect more negatively the results than the night conditions.

As for the use of hierarchical localization, it may result interesting for high levels of compression and just selecting one cluster as candidate may be enough for most cases.

Future works will include the study of other methods to compress the models and the study of other disadvantageous issues which may be presented in real operating conditions, such as occlusions, changes of furniture, etc.

ACKNOWLEDGMENTS

The authors declare that there are no competing interests regarding the publication of this paper.

REFERENCES

[1] D. Valiente, A. Gil, Ó. Reinoso, M. Juliá, and M. Holloway, "Improved omnidirectional odometry for a view-based mapping approach," *Sensors*, vol. 17, no. 2, p. 325, 2017.

[2] L. B. Marinho, J. S. Almeida, J. W. M. Souza, V. H. C. Albuquerque, and P. P. R. Filho, "A novel mobile robot localization approach based on topological maps using classification with reject option in omnidirectional images," *Expert Syst. Appl.*, vol. 72, pp. 1–17, Apr. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417416306790

[3] M. Faessler, F. Fontana, C. Forster, E. Mueggler, M. Pizzoli, and D. Scaramuzza, "Autonomous, vision-based flight and live dense 3D mapping with a quadrotor micro aerial vehicle," *J. Field Robot.*, vol. 33, no. 4, pp. 431–450, 2016.

[4] Y. Berenguer, L. Payá, M. Ballesta, and O. Reinoso, "Position estimation and local mapping using omnidirectional images and global appearance descriptors," *Sensors*, vol. 15, no. 10, pp. 26368–26395, 2015. [Online]. Available: http://www.mdpi.com/1424-8220/15/10/26368

[5] A. Angeli, S. Doncieux, J.-A. Meyer, and D. Filliat, "Visual topological SLAM and global localization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2009, pp. 2029–2034.

[6] A. C. Murillo, J. J. Guerrero, and C. Sagues, "Surf features for efficient robot localization with omnidirectional images," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3901–3907.

[7] E. Garcia-Fidalgo and A. Ortiz, "Vision-based topological mapping and localization by means of local invariant features and map refinement," *Robotica*, vol. 33, no. 7, pp. 1446–1470, 2015.

[8] E. Menegatti, T. Maeda, and H. Ishiguro, "Image-based memory for robot navigation using properties of omnidirectional images," *Robot. Auton. Syst.*, vol. 47, no. 4, pp. 251–267, 2004.

[9] M. Liu, D. Scaramuzza, C. Pradalier, R. Siegwart, and Q. Chen, "Scene recognition with omnidirectional vision for topological map using lightweight adaptive descriptors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2009, pp. 116–121.

[10] H. Korrapati and Y. Mezouar, "Multi-resolution map building and loop closure with omnidirectional images," *Auto. Robots*, vol. 41, no. 4, pp. 967–987, Mar. 2016.

[11] L. Gerstmayr-Hillen, F. Röben, M. Krzykawski, S. Kreft, D. Venjakob, and R. Möller, "Dense topological maps and partial pose estimation for visual control of an autonomous cleaning robot," *Robot. Auton. Syst.*, vol. 61, no. 5, pp. 497–516, 2013.

[12] H. Korrapati and Y. Mezouar, "Vision-based sparse topological mapping," *Robot. Auton. Syst.*, vol. 62, no. 9, pp. 1259–1270, 2014.

[13] F. Amigoni *et al.*, "A standard for map data representation: IEEE 1873–2015 facilitates interoperability between robots," *IEEE Robot. Autom. Mag.*, vol. 25, no. 1, pp. 65–76, Mar. 2018.

[14] A. Štimec, M. Jogan, and A. Leonardis, "Unsupervised learning of a hierarchy of topological maps using omnidirectional images," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 22, no. 4, pp. 639–665, 2007. doi: 10.1142/S0218001408006430.

[15] E. Garcia-Fidalgo and A. Ortiz, "Vision-based topological mapping and localization methods: A survey," *Robot. Auton. Syst.*, vol. 64, pp. 1–20, Feb. 2015. doi: 10.1016/j.robot.2014.11.009.

[16] S. P. P. da Silva, R. V. M. da Nóbrega, A. G. Medeiros, L. B. Marinho, J. S. Almeida, and P. P. R. Filho, "Localization of mobile robots with topological maps and classification with reject option using convolutional neural networks in omnidirectional images," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[17] A. Rituerto, L. Puig, and J. J. Guerrero, "Visual SLAM with an omnidirectional camera," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 348–351.

[18] S. Cebollada, L. Payá, W. Mayol, and O. Reinoso, "Evaluation of clustering methods in compression of topological models and visual place recognition using global appearance descriptors," *Appl. Sci.*, vol. 9, no. 3, p. 377, 2019.

[19] R. González and R. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2008.

[20] L. Payá, W. Mayol, S. Cebollada, and O. Reinoso, "Compression of topological models and localization using the global appearance of visual information," in *Proc. ICRA*, 2017, pp. 5630–5637.

[21] L. Payá, O. Reinoso, Y. Berenguer, and D. Úbeda, "Using omnidirectional vision to create a model of the environment: A comparative evaluation of global-appearance descriptors," *J. Sensors*, vol. 2016, Feb. 2016, Art. no. 1209507. doi: 10.1155/2016/1209507.

[22] L. Payá, A. Peidró, F. Amorós, D. Valiente, and O. Reinoso, "Modeling environments hierarchically with omnidirectional imaging and global-appearance descriptors," *Remote Sens.*, vol. 10, no. 4, p. 522, 2018.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, vol. 2, Jun. 2005, pp. 886–893.

[24] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *Comput. Vis. Image Understand.*, vol. 78, no. 1, pp. 99–118, 2000.

[25] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Prog. Brain Res.*, vol. 155, pp. 23–36, 2006.

[26] A. Murillo, G. Singh, J. Košecká, and J. J. Guerrero, "Localization in urban environments using a panoramic gist descriptor," *IEEE Trans. Robot.*, vol. 29, no. 1, pp. 146–160, Feb. 2013.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[28] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[29] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. *Places-CNN Model From MIT*. Accessed: Feb. 28, 2019 [Online]. Available: https://github.com/BVLC/caffe/wiki/Model-Zoo#places-cnn-model-from-mit

[30] M. Mancini, S. R. Bulò, E. Ricci, and B. Caputo, "Learning deep NBNN representations for robust place categorization," *IEEE Robot. Autom. Lett.*, vol. 2, no. 3, pp. 1794–1801, Jul. 2017.

[31] Y. Xu, M. Sun, Z. Cao, J. Liang, and T. Li, "Multi-object tracking for mobile navigation in outdoor with embedded tracker," in *Proc. 7th Int. Conf. Natural Comput. (ICNC)*, vol. 3, 2011, pp. 1739–1743.

[32] S. Theodoridis and K. Koutroumbas, "Pattern recognition and neural networks," in *Proc. Mach. Learn. Appl.*, 2001, pp. 169–195.

[33] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[34] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2001, pp. 849–856.

[35] A. Pronobis and B. Caputo, "COLD: COSY localization database," *Int. J. Robot. Res.*, vol. 28, no. 5, pp. 588–594, May 2009. [Online]. Available: http://www.pronobis.pro/publications/pronobis2009ijrr

SERGIO CEBOLLADA received the M.Eng. degree in telecommunication engineering from Miguel Hernández University, in 2014, where he is currently pursuing Ph.D. degree. His research interests include omnidirectional vision and global appearance algorithms, map building and localization of mobile robots, and deep learning. He received the Ph.D. Candidate Scholarship Supported by Valencian Government (ACIF/2017/146), in 2017.

LUIS PAYÁ received the M.Eng. degree in industrial engineering in Spain, in 2002, and the Ph.D. degree in industrial technologies in Spain, in 2014. He is currently an Associate Professor with the Department of Systems Engineering and Automation, Miguel Hernández University, Spain. He teaches some subjects related to the fields of automatic control, electronics, and robotics. He has authored several books, papers, and communications in the cited topics. His current research interests include omnidirectional vision and global appearance algorithms, topological map building and localization of mobile robots, and also implementation and testing of remote laboratories.

VICENTE ROMÁN received the Degree in electronics and automation engineering, in 2016 and the M.Sc. degree in robotics from Miguel Hernandez University, Elche, Spain, in 2017, where he is currently pursuing the Ph.D. degree. He teaches some subjects related to control, robotics, and computer vision with the Department of Systems Engineering and Automation, UMH. His current research interests include mobile robotics, omnidirectional vision, and global appearance algorithms.

OSCAR REINOSO received the Degree in industrial engineering and the Ph.D. degree from the Polytechnic University of Madrid, in 1991 and 1996, respectively. From 1994 to 1997, he was with the Research and Development Department, Protos Desarrollo in a visual inspection system. Since 1997, he has been with Miguel Hernández University, as a Professor in control, robotics, and computer vision. He has authored several books, papers, and communications in the cited topics. His research interests include robotics, teleoperated robots, climbing robots, visual serving, and visual inspection systems. He is a member of the CEA-IFAC.