View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Explore Bristol Research

Paya Castello, L., Mayol-Cuevas, W., Cebollada, S., & Reinoso, O. (2017). Compression of topological models and localization using the global appearance of visual information. In 2017 IEEE International Conference on Robotics and Automation (ICRA 2017). 9781509046348: Institute of Electrical and Electronics Engineers (IEEE).

Peer reviewed version

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) will be available online via IEEE. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms.html

Compression of topological models and localization using the global appearance of visual information*

Luis Payá[1], Walterio Mayol[2], Sergio Cebollada[3] Oscar Reinoso[4]

Abstract— **In this work, a clustering approach to obtain compact topological models of an environment is developed and evaluated. The usefulness of these models is tested by studying their utility to solve the robot localization problem subsequently. Omnidirectional visual information and global appearance descriptors are used both to create and compress the models and to estimate the position of the robot. Comparing to the methods based on the extraction and description of landmarks, global appearance approaches permit building models that can be handled and interpreted more intuitively and using relatively straightforward algorithms to estimate the position of the robot. The proposed algorithms are tested with a set of panoramic images captured with a catadioptric vision sensor in a large environment under real working conditions. The results show that it is possible to compress substantially the visual information contained in topological models to arrive to a balance between the computational cost and the accuracy of the localization process.**

I. INTRODUCTION

Currently, omnidirectional imaging has become a popular option in the development of mapping and localization tasks, thanks to the large quantity of information the images offer as they cover a field of view of 360° around the robot. However, the high dimensionality of the data requires a processing step to extract relevant information from scenes. This information must be useful to create a compact model that permits a robust and computationally efficient localization. The use of omnidirectional vision would also permit developing localization algorithms which are invariant to the orientation of the robot when its movement is contained in the ground plane, which would increase the scope of these algorithms.

Depending on how the most relevant information from the images is extracted and represented, visual map building and localization has been approached using mainly two main frameworks. The first one is based on the detection, description and tracking of some relevant local features along a set of scenes [1], [2]. The second method consists in working with each scene as a whole, building a unique descriptor per image that contains information on its global appearance [3], [4], [5]. These methods usually lead to more intuitive representations of the environment and relatively straightforward localization algorithms, based mainly on the pairwise comparison between descriptors. However, due to their lack of metric information, they have been used traditionally to build topological representations of the environment [6], which are accurate enough for many applications. When necessary, they can be combined with metric data into a hybrid map where the information is arranged into several layers, from a topological high-level layer that permits a rough but quick localization to some metric or topometric low-level layers to refine the position, if necessary [7]. Some authors have proposed data compression strategies to build efficient maps based on local features [8], [9], [10].

To create a visual model, initially, a set of images captured from several points of view of the environment to map are usually available. Among the compressing methods, clustering algorithms can be used to compact the information on this model to create a high level map, which would group this set of images into several clusters containing visually similar scenes and represent each cluster with a representative instance. Some research has made use of such algorithms previously to create visual maps, using local features, such as Valgren *et al.* [11] and Stimec *et al.* [12]. Ideally, using only visual information, a clustering algorithm should group images that have been captured in geometrically close points. However, many indoor environments are prone to *visual aliasing* and images that have been captured far away can be quite similar, which would lead to errors in the model.

The main objective of this work consists in carrying out an exhaustive comparative evaluation of several image descriptors to create compact models of an unknown environment and to solve the localization problem using these models. During the implementation of hybrid mapping algorithms, the high level maps must be useful to estimate roughly the position of the robot with a low computational cost. The results of the present work may be helpful in this field as they will permit choosing the best description method and tuning correctly the main parameters in such a way that a compromise between computational cost and accuracy are reached. To build and compress the models only visual information is considered. We will focus on the use of global appearance descriptors, since local descriptors have been considered in previous research. It leads to purely topological models that must be able to cope with the visual aliasing phenomenon. A challenging set of images captured in an indoor environment, including regions with similar appearance and changes in lighting conditions during the capture process is used to test the algorithms developed.

*This work was supported by the Spanish Government through the projects DPI2013-41557-P and DPI2016-78361-R and by Miguel Hernández University, through the grant AC1006VM.

[1]L. Payá is with Department of Engineering Systems and Automation, Miguel Hernández University, Elche, Spain lpaya@umh.es

[2]W. Mayol is with the Department of Computer Science, University of Bristol, Bristol, United Kingdom wmayo@cs.bris.ac.uk

[3]O. Reinoso is with Department of Engineering Systems and Automation, Miguel Hernández University, Elche, Spain o.reinoso@umh.es

[3]S. Cebollada is with Department of Engineering Systems and Automation, Miguel Hernández University, Elche, Spain sergio.cebollada@umh.es

This work continues the research started in [13], where a comparative evaluation between some global description methods is carried out to obtain a high-level model and some low-level topological maps of an environment. Now, this evaluation is extended to the compacting and localization problems. The remainder of the paper is structured as follows. Section II makes a brief outline of the global appearance approaches that will be tested along the paper. After that, section III presents the clustering approaches used to create compact models and section IV tests the validity of these models to solve the localization problem. At last, a final discussion is carried out in section V.

II. DESCRIBING THE GLOBAL APPEARANCE OF A SET OF SCENES

This section outlines some methods to describe the global appearance of a set of scenes. Three families of methods are proposed to be evaluated along the paper: methods based on the Discrete Fourier Transform (subsection II-A), on histograms of orientation gradients (subsection II-B) and on the gist of the scenes (subsection II-C). Also, since changing lighting conditions may have a pronounced effect on the global appearance of the scenes, the use of homomorphic filtering [14] will be tested to cope with this problem.

A complete description of these methods can be found in [13]. In all cases, the starting point is a panoramic scene $im(x, y) \in \mathbb{R}^{N_x \times N_y}$ and after using any of these methods the result is a global appearance descriptor $\vec{d} \in \mathbb{R}^{l \times 1}$

A. Fourier Signature

The formulation of the Discrete Fourier Transform (DFT) we use along the paper is known as Fourier Signature (FS), and was used initially by Menegatti *et al.* [15] to create a visual memory of an unknown environment. It is defined as the matrix composed of the 1-D DFT of each row in the original image. The FS of a panoramic image $im(x, y) \in \mathbb{R}^{N_x \times N_y}$ is a new matrix $IM(u, y) \in \mathbb{C}^{N_x \times N_y}$ (u is the frequency variable, measured in cycles/pixel). The main information is concentrated in the low frequency components and the high frequency ones tend to be more contaminated by the eventual presence of noise in the scene, so only the k_1 first columns can be retained, having a compression effect. The new matrix $IM(u, y) \in \mathbb{C}^{N_x \times k_1}$ can be decomposed into two real matrices, one containing the magnitudes $A(u, y)$ and the other with the arguments, $\Phi(u, y)$, both of them with N_x rows and k_1 columns. $A(u, y)$ is invariant against changes of the robot orientation on the ground plane and can be arranged to compose a global appearance descriptor $\vec{d} \in \mathbb{R}^{N_x \cdot k_1 \times 1}$ of the original panoramic image $im(x, y)$.

B. Histogram of Oriented Gradients

Initially described by Dalal and Triggs [16] to solve people detection tasks, the Histogram of Oriented Gradients (HOG) considers the gradient orientation in local areas of a scene to build a descriptor. The method stands out by its simplicity and efficiency in object recognition tasks. The use of HOG in robot mapping and localization is somewhat sparse and usually limited to small and controlled environments [17]. The version of HOG we consider is described in [18], where the method is redefined with the goal of obtaining a unique and rotationally invariant global appearance descriptor per scene. Basically it consists in dividing the panoramic image in a set of k_2 horizontal cells and compiling a histogram per cell with b bins each, that reflect the gradient orientation of the pixels within this cell. This set of histograms compose the final descriptor $\vec{d} \in \mathbb{R}^{b \cdot k_2 \times 1}$.

C. Gist of a scene

Inspired by the human process to recognize scenes, Oliva *et al.* [19] developed the *gist* concept with the idea of creating a low-dimension global scene descriptor. More recently, it has been used often together with the *prominence* concept, which refers to the properties that make a pixel to stand out with respect to its neighbours. Siagian *et al.* [20] tried to establish a synergy between the two concepts and they designed a unique descriptor that takes both into account. This descriptor is built using the intensity, orientation and color information. The experience with this kind of descriptors in mobile robots applications is quite limited. For example, Chang *et al.* [21] present a localization and navigation system based on *gist* and prominence and Murillo *et al.* [22] make use of *gist* descriptors in a localization problem. However, they obtain these descriptors using specific regions in a set of panoramic images. The implementation of *gist* used in this work is described in [18], and offers a rotationally invariant version of *gist* when applied to panoramic images. Basically, the descriptor is built from orientation information, obtained after applying several Gabor filters with m_1 different orientations to the original image in m_2 resolution levels. The information is then reduced by grouping the pixels of every resulting image into k_3 horizontal blocks. The result is a descriptor $\vec{d} \in \mathbb{R}^{m_1 \cdot m_2 \cdot k_3 \times 1}$.

III. COMPACTING VISUAL MODELS USING A CLUSTERING APPROACH

This section focuses on the creation of the topological model and the compression of this model. The starting point is a set of panoramic images $I = \{im_1, im_2, \ldots, im_n\}$ captured from several points of view, covering the whole environment to model. These images may be optionally filtered with a homomorphic filter and then each image is globally described using any of the methods described in section II. As a result, the original model will be composed of a set of descriptors $\mathbf{D} = \{\vec{d_1}, \vec{d_2}, \ldots, \vec{d_n}\}$. The coordinates of the capture points of each images are also known $\mathbf{P} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. During the experiments only visual information will be used to build the model and estimate the position of the robot, and these coordinates will only be used as *ground truth* to assess the performance of the algorithm.

A. Compacting the model

A functional map should permit carrying out the localization process with a reasonable accuracy and computational cost. In this work, a clustering approach will be used to compress the information of the original model. The original data set $\mathbf{D} = \{\vec{d_1}, \vec{d_2}, \ldots, \vec{d_n}\}$ will be divided into m clusters $\{C_1, C_2, \ldots, C_m\}$ such that:

$$C_i \neq \emptyset, i = 1, \ldots, m$$
$$\bigcup_{i=1}^{m} C_i = \mathbf{D} \qquad (1)$$
$$C_i \cap C_j = \emptyset, i \neq j, i, j = 1, \ldots m.$$

After the clustering process, each cluster will be reduced to a representative descriptor so the compact model will consist of a set of representatives $\mathbf{R} = \{\vec{r_1}, \vec{r_2}, \ldots, \vec{r_m}\}$.

To compress the initial model, the original set of descriptors $\mathbf{D}$ is used as input data to the clustering algorithm creating thus clusters containing images whose visual appearance is similar. Ideally those images with similar appearances should correspond to images captured in geometrically near positions. However, often this is not true due to the phenomenon of visual aliasing, which is frequent in structured indoor environments. This fact, together with the high dimensionality of the data, would make it unfeasible the use of classical clustering algorithms, based on the optimization of a function, such as k-means or hierarchical algorithms. Instead, spectral clustering has proved to cluster successfully such high dimensional data [23], including visual information [11], [12]. The implementation used in this paper was developed by Ng *et al* [24].

The spectral clustering algorithms take into account the mutual similarity among all the instances. This is why they have proved to be more effective than traditional methods, which only consider the similarity between each instance and the m representatives. The algorithm starts obtaining the mutual similarity between instances $\mathbf{S}_{ij}$ to build the matrix $\mathbf{S}$ (eq. 2). Algorithm 1 shows the complete process.

$$\mathbf{S}_{ij} = e^{-\frac{|\vec{d_i} - \vec{d_j}|^2}{2\sigma^2}} \qquad (2)$$

Algorithm 1 Spectral Clustering Algorithm

```
Input: Similarity matrix $\mathbf{S}$, number of clusters $m$
Output: Set of clusters $C_1, C_2, \ldots, C_m$
$\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{S}_{ij}$, {$\mathbf{D}$ diagonal matrix}
$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{1/2}$ {Laplacian matrix}
$\mathbf{U} \leftarrow m$ main eigenvectors of $\mathbf{L}$ in columns
$\mathbf{T} \leftarrow$ normalized rows of $\mathbf{U}$
Clusters $A_1, \ldots, A_m \leftarrow$ k-means clustering considering
    as instances the rows of $\mathbf{T}, \vec{t_i}, i = 1, \ldots n$
$C_1, \ldots, C_m$ such that $C_i = \vec{d_j} | \vec{t_j} \in A_i$
```

When the number of instances n or their dimension l is very high, calculating the m main eigenvectors of the Laplacian matrix can be computationally expensive. One possible solution to this problem consists in canceling some of the components of the similarity matrix, to obtain a sparse matrix, and using sparse matrices methods to obtain the required eigenvectors [23]. To do that, in the matrix $\mathbf{S}$ only the components $\mathbf{S}_{ij}$ such that j is among the t nearest neighbours of i or vice versa, being t a low number, are retained in this work. After that, the Lanczos/Arnoldi factorization is used to obtain the first m eigenvectors of the Laplacian matrix $\mathbf{L}$.

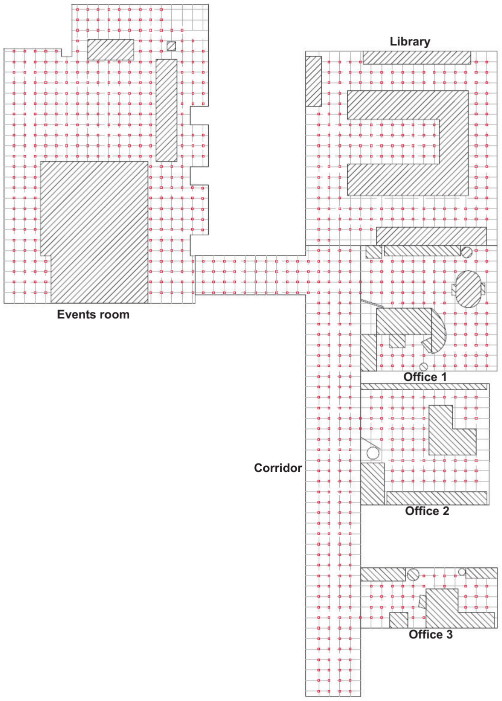


Fig. 1. Bird's eye view of the capture points of the training set of images. The size of the grid is 40 × 40 cm.

Once the clusters have been created, the representatives are obtained as the average visual descriptor of all the descriptors included in each cluster. These representatives are the set $\{\vec{r_1}, \ldots, \vec{r_m}\}$.

B. Experiments

To carry out the experiments, a complete database captured by ourselves is used. This database is publily available [25]. It was captured using a catadioptric vision system composed of a *Eizoh Wide 70* hyperbolic mirror mounted over an *Imaging Source DFK 21BF04* camera with their axes aligned, and covers the whole floor of a building of Miguel Hernández University (Spain), including 6 different rooms. This database includes two sets of images. The first one (training set) consists of 872 panoramic 64×256 images captured on a dense 40×40 cm grid of points. It will be used to build the visual model of the environment. The second set (test set) consists of 546 images captured in all the rooms, in some half-way points among the grid positions and with different orientations, times of day and changes in the position of some objects. This set will be used during the localization process, to test the usefulness of the previously built model. Fig. 1 shows a bird's eye view of the environment and the capture points of the training images.

Fig. 2 shows, the results of a sample clustering experiment

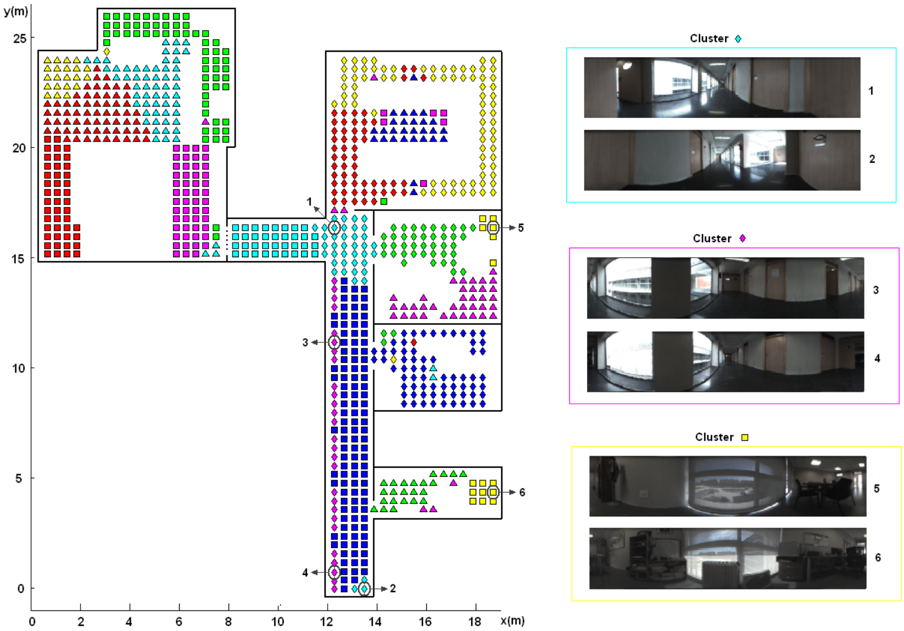


Fig. 2. Results of a sample clustering experiment considering $m = 18$ clusters and gist descriptor with $k_3 = 16$, $m_1 = 32$, $m_2 = 2$ and some sample panoramic images belonging to the set 1.

applied to this dataset, according to algorithm 1. The figure presents a bird's eye view of the capture points of the set 1, showing with different colors and shapes the images belonging to each cluster. To obtain this figure we consider $m = 18$ clusters and the gist descriptor with $k_3 = 16$ cells, $m_1 = 32$ Gabor masks and $m_2 = 2$ levels. Since only visual information is used to create the clusters, some of them tend to be separated among several rooms or are relatively little compact, due to the effect of visual aliasing. This effect is clearly shown through several sample panoramic images belonging to the set 1 in fig. 2.

To assess the performance of each description method, some parameters are measured to study (a) the compactness of the clusters (i.e. if they really group images which have been captured in geometrically close points) and (b) if the number of instances is balanced among clusters. First, the geometrical compactness of each cluster is measured through its moment of inertia and its silhouette. After a clustering process, the average moment of inertia is calculated as:

$$\mathbf{M} = \sum_{i=1}^{m} \left(\frac{\sum_{j=1}^{n_i} \left[dist\left(\vec{d}_{ij}, \vec{c}_i\right)\right]^2}{n_i} \right) \tag{3}$$

where the cluster C_i is composed of the descriptors $\{\vec{d}_{i1}, \vec{d}_{i2}, \ldots, \vec{d}_{in_i}\}$, n_i is the number of images in cluster C_i and $dist\left(\vec{d}_{ij}, \vec{c}_i\right)$ is the Euclidean distance between the capture point of the image j of the cluster i and the position of this cluster representative $\vec{c}_i$. The lower is $\mathbf{M}$ the more compact are the clusters.

The silhouette is a classical way of interpretation and validation of clusters that gives us an idea of the degree of similarity between each entity and the other entities of the same cluster, comparing it with the entities in the other clusters. The higher is S, the more similar is each entity to its own cluster and the more different to the entities on the other clusters. In this work, the silhouette is used to evaluate the compactness of the clusters. This way, instead of using the similarity in feature space, world coordinates are used. The average silhouette is calculated as:

$$\mathbf{S} = \frac{\sum_{k=1}^{n} s_k}{n} \tag{4}$$

where n is the number of instances to cluster and s_k is the silhouette of each instance $\vec{d}_k$, $s_k = \frac{b_k - a_k}{\max(a_k, b_k)}$. a_k is the average distance between the capture point of $\vec{d}_k$ and the capture point of the other entities contained in the same cluster and b_k is the minimum average distance between the capture point of $\vec{d}_k$ and the capture points of the entities contained in the other clusters.

Second, their balance is measured through the standard deviation of the number of instances per cluster $\mathbf{D}$. The lower is $\mathbf{D}$, the more balanced is the number of instances of all the clusters.

Fig. 3 shows the results of both clustering methods using FS as descriptor (top row), depending on the configuration of the parameter k_1 and the *gist* descriptor (bottom row), depending on the configuration of the parameter k_3. The average silhouette $\mathbf{S}$, moment of inertia $\mathbf{M}$ and deviation $\mathbf{D}$ vs. the number of clusters m is depicted in both cases. For comparative purposes, the same scale is used in the vertical axis of each pair of graphical representations. The HOG descriptor has shown to be unable to create clusters that tend to group images captured from near positions, as fig 4 shows. As far as gist is concerned, high values of k_3 produce a changing behaviour, depending on m. A very low value

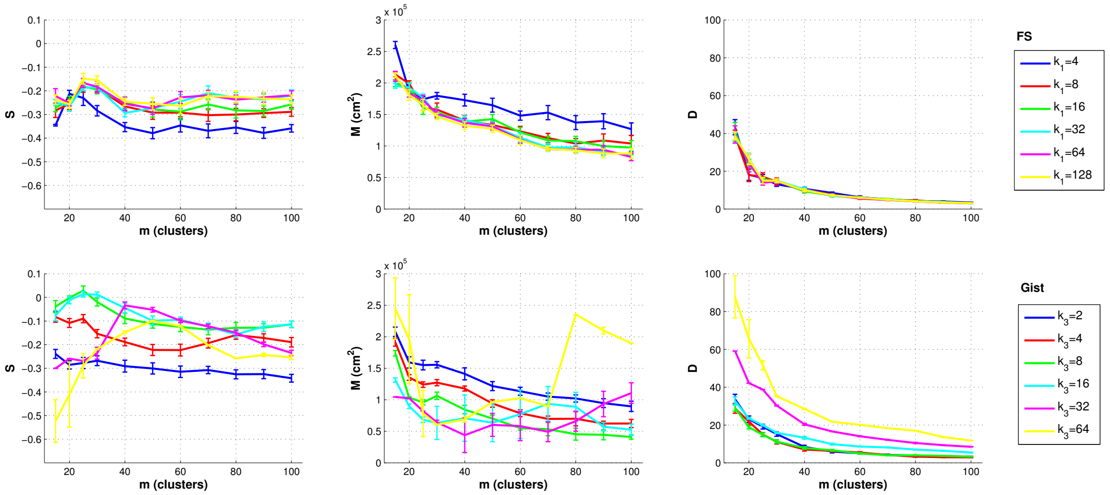


Fig. 3. Results of the clustering process: average silhouette, average moment of inertia and average deviation vs. number of clusters, when using FS (top row) and (b) gist (bottom row)

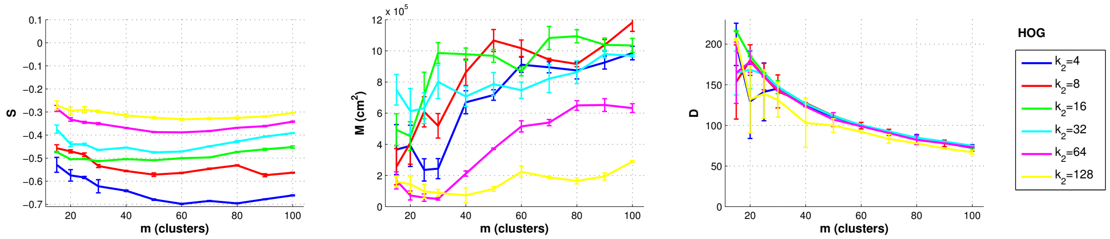


Fig. 4. Results of the clustering process: average silhouette, average moment of inertia and average deviation vs. number of clusters, when using HOG

for k_3 tends to produce low silhouettes and high moments, independently on the number of clusters. This way, the best choice could be an intermediate value for k_3. In general, $k_3 = [8, 16]$ produces relatively good results in general, and the best values of silhouette and moment are achieved with $k_3 = 32$ and an intermediate number of clusters. If we compare the results of gist with FS, we can arrive to the conclusion that gist outperforms FS, in general. The silhouette of FS tends to be lower and the moment higher. As far as the dispersion in the number of entities per cluster, in general FS presents a slightly higher dispersion comparing to gist except when $k_3 = [32, 64]$.

IV. SOLVING THE LOCALIZATION PROBLEM USING THE COMPACT TOPOLOGICAL MAPS

In the previous section, gist has proved to be an efficient descriptor to create a compact model. Once built, the utility of this model to solve the localization problem can be evaluated. In this section, a comparative study of the performance of the proposed description methods during this localization step is carried out.

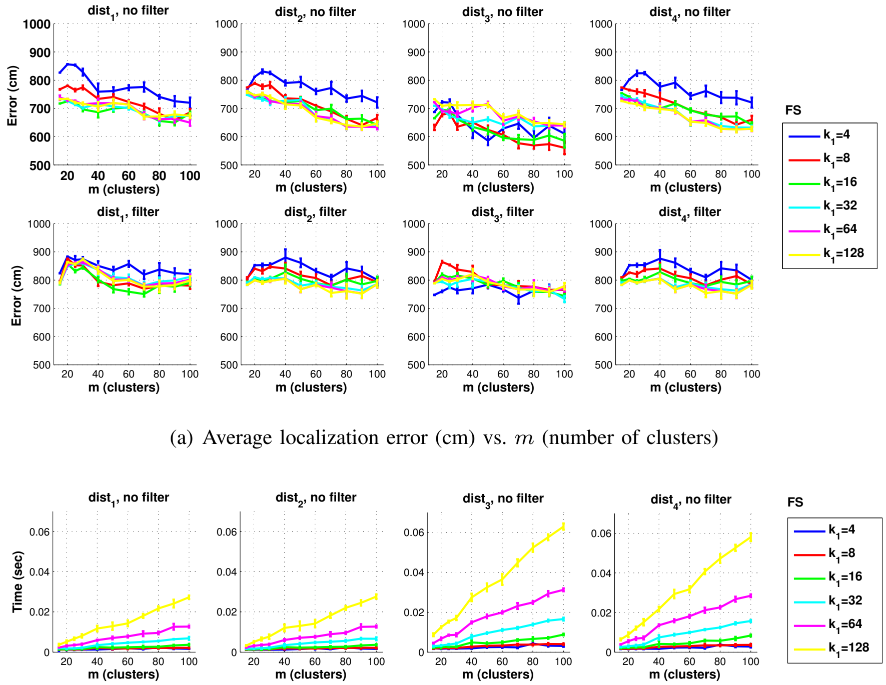
A. Localization process

After the clustering process, the compact topological model consists of a set of cluster representatives $\{\vec{r}_1, \ldots, \vec{r}_m\}$. The coordinates of the center of each cluster are also known $\{(x, y)_1, \ldots, (x, y)_m\}$. However, these coordinates will only be used to test the accuracy of the localization method (ground truth), but not to estimate the position of the robot. A purely visual approach will be used with this aim.

The localization is addressed in an absolute fashion, assuming no information on the previous position of the robot is known: the robot captures a new test image im_t, filters it, describes it to obtain $\vec{d_t}$ and calculates the distance between it and each representative, obtaining the distances vector $\vec{l_t} = \{l_{t1}, \ldots, l_{tm}\}$ where $l_{tj} = dist\{\vec{d_t}, \vec{r_j}\}$ according to any distance measure. The node that presents the minimum distance $d_t^{nn}|t = arg\min_j l_{tj}$ is considered the corresponding position of the robot. To estimate the accuracy of this correspondence, the geometric distance between the capture point of the test image and the center of the corresponding cluster is calculated. Also, the computational time of the localization process will be evaluated, with the objective of arriving to a balance between degree or compression and accuracy in relocalization. The experiments have been carried out through Matlab programming.

B. Experiments

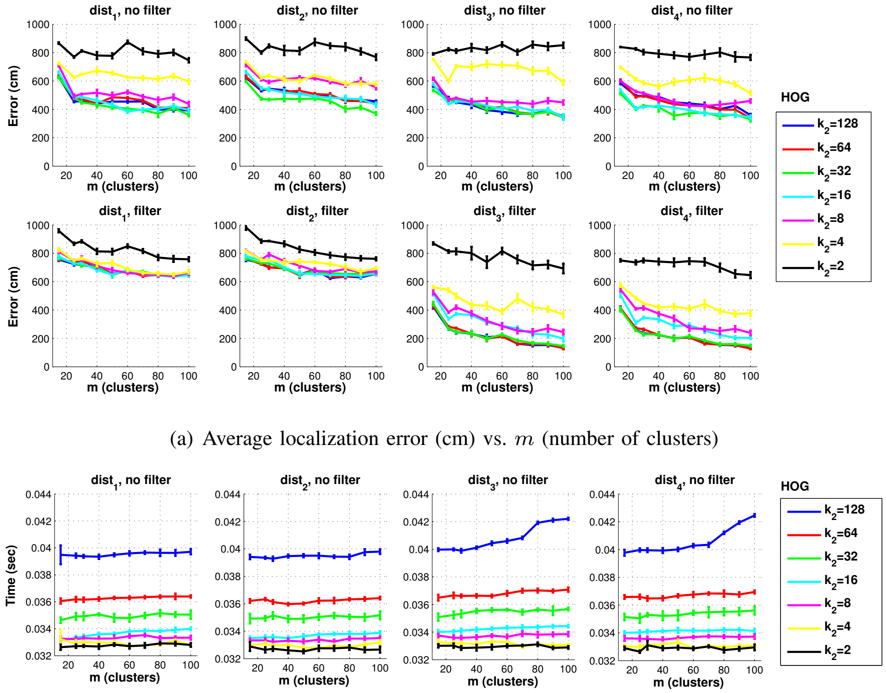
The starting point of the localization experiment is a compact model. To create this model, different numbers of clusters will be considered, from $m = 15$ to $m = 100$. Since the clustering results may change from experiment to experiment (due to the random initialization of the representatives



(a) Average localization error (cm) vs. m (number of clusters)

(b) Average computation time (sec) vs. m (number of clusters)

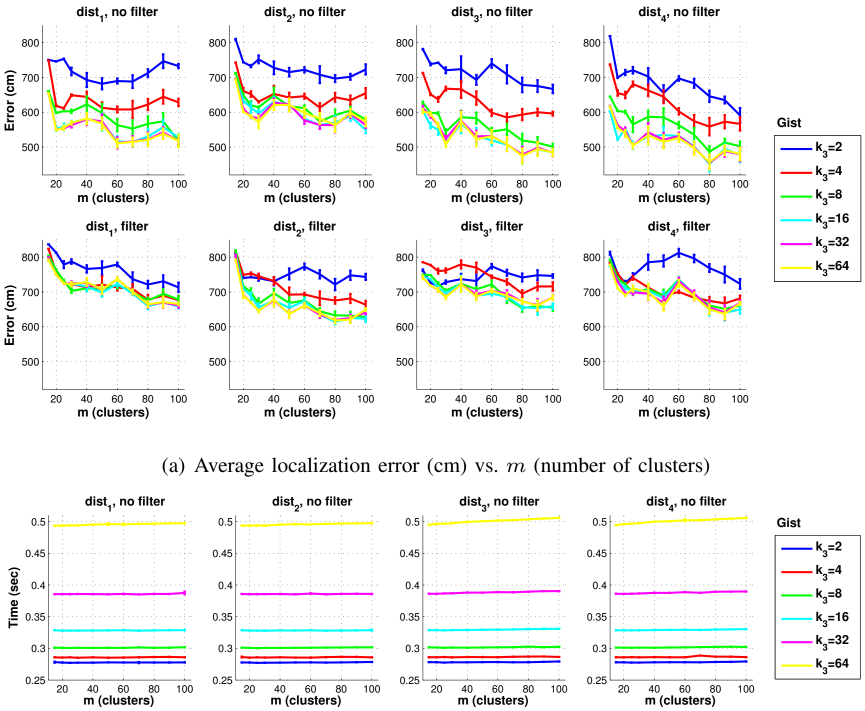
Fig. 5. Results of the localization process when FS is used to describe both the test images and the clusters' representatives, depending on k_1, the distance measure and the use of homomorphic filter.



(a) Average localization error (cm) vs. m (number of clusters)

(b) Average computation time (sec) vs. m (number of clusters)

Fig. 6. Results of the localization process when HOG is used to describe both the test images and the clusters' representatives, depending on k_2, the distance measure and the use of homomorphic filter



(a) Average localization error (cm) vs. m (number of clusters)

(b) Average computation time (sec) vs. m (number of clusters)

Fig. 7. Results of the localization process when gist is used to describe both the test images and the clusters' representatives, depending on k_3, the distance measure and the use of homomorphic filter.

in the k-means algorithm run in the step 5 of the algorithm 1), a total of 100 models have been created, for each value of m. The localization experiments will be carried out with respect to all these models and the average results will be shown in the figures. To create the clusters, the gist descriptor with $k_3 = 16$ has been chosen, as this configuration has offered relatively good results in the previous section.

To carry out the experiments each image of the set 2 is considered as a test image. These test images were captured from different positions of those of the map, in different times of day (including severe changes in lighting conditions), with the presence of people and changes in the position of doors and other objects. This way, the experiment is carried out under real localization conditions.

Each test image is compared with the representatives of the compact map and the most similar cluster is retained. In the experiments, 4 distance measures are considered to carry out this comparison: $dist_1$ is the Manhattan distance, $dist_2$ is the Euclidean distance, $dist_3$ is the correlation distance and $dist_4$ is the cosine distance. Fig. 5 shows the results obtained when FS is used as description method, fig. 6 with HOG and, finally, the results of gist are shown in fig. 7. In all these figures both the average localization error (cm) and the computational time (sec) vs. the number of clusters m are shown, and the influence of the type of distance is also assessed. Finally, the effect of using optionally homomorphic filtering is also shown in the localization error figures. In the case of the computational time, the use of filter adds a constant average time equal to 0.02 sec per test image.

In general, when the number of clusters increases, the computational time also increases and the localization error decreases. This is an expected result since when the number of clusters is low, the distance between cluster representatives is larger and thus, the localization error with respect to these representatives tends to be higher. This way, for each description method, the number of clusters may be tuned to reach a compromise between error and time. We must take it into account that some association errors may happen during the localization process, and this effect will also be reflected in the figures. The lower is m, the more association errors are expected to happen.

Making an individual analysis of the relation between the type of descriptor and the localization error, the following conclusions can be reached. First, FS presents a relatively high localization error and the use of homomorphic filter does not improve the situation. The minimum localization error is obtained with $dist_3$ (correlation) and without filter. This error takes values between 5.7 m and 7 m depending on the size of the descriptor and the number of clusters. Second, in the case of HOG the use of filter clearly improves the results when using $dist_3$ and $dist_4$. In this case, the localization error takes values between 1.8 m and 5.8 m depending on the number of clusters, when the dimension of the descriptor takes intermediate or high values. Third, gist does not improve the results provided by HOG. In the case of gist, the use of filter worsens the localization results, and the best results are obtained with $dist_3$ and $dist_4$ and a high number of components in the descriptor. About the computational cost, the FS tends to need less time to solve the localization problem, followed by HOG and gist, whose computational cost is, in general, one order higher than the cost of HOG.

As a final conclusion, as far as the localization error is concerned, the optimal values are obtained with HOG with homomorphic filter, considering high values of k_2. This error falls under 2 m when the number of clusters is $m > 50$, which is a relatively low value taking into account the total area of the mapped environment.

V. CONCLUSION AND FUTURE WORKS

In this work, the problem of creating compact topological maps has been addressed. A set of 872 panoramic images has been used to model a large indoor environment, and a clustering approach has been implemented to compress the information in this model and reduce it to a lower number of instances. We have considered between 15 and 100 instances in the compact model, what supposes having reduced the number of instances to between 1.7% and 11.5% of the complete initial model. Once created the compact models, their utility to solve the robot relocalization task has been tested.

The problem has been approached using the global appearance of panoramic scenes both to compact the model and to solve the localization problem. A comparative evaluation between three global appearance descriptors has been carried out and their performance has been tested depending, basically, on the size of the descriptor.

The work has shown how it is possible to compress substantially the visual information in the original model (thus reducing the computational cost of the localization process) while keeping a relatively good localization error. On the one hand, the gist descriptor has proved to be the best choice to compress the model, through the clustering approach implemented, since it has produced the most compact and balanced clusters. It has an ability to group images that have been captured in close points despite the visual aliasing. On the other hand, once the clusters have been created, the use of HOG to describe both the cluster representatives and the test image is the choice that has presented the best localization results, used jointly with the homomorphic filter and the correlation distance.

Future works will include, on the one hand, the study of other methods to compress the models and a comparative evaluation with methods based on local features or landmarks and, on the other hand, the adaptation of the method to be used in lifelong map updating, exploring the use of an incremental clustering algorithm with this aim, to be used in long term operation.

REFERENCES

[1] C. Valgren and A. Lilienthal, "SIFT, SURF & seasons: Appearance-based long-term localization in outdoor environments," *Robotics and Autonomous Systems*, vol. 58, pp. 149–156, 2010.

[2] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," pp. 2564–2571, 2011.

[3] B. Krose, R. Bunschoten, S. Hagen, B. Terwijn, and N. Vlassis, "Visual homing in enviroments with anisotropic landmark distrubution," 2007.

[4] A. Leonardis and H. Bischof, "Robust recognition using eigenimages," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 99–118, 2000.

[5] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," pp. 1023–1029, 2000.

[6] E. Garcia-Fidalgo and A. Ortiz, "Vision-based topological mapping and localization methods: A survey," *Robotics and Autonomous Systems*, vol. 64, pp. 1–20, 2015.

[7] I. Kostavelis, K. Charalampous, A. Gasteratos, and J. Tsotsos, "Robot navigation via spatial and temporal coherent semantic maps," *Engineering Applications of Artificial Intelligence*, vol. 48, pp. 173–187, 2016.

[8] L. Contreras and W. Mayol-Cuevas, "Trajectory-driven point cloud compression techniques for visual SLAM," pp. 133–140, 2015.

[9] S. Rady, A. Wagner, and E. Badreddin, "Building efficient topological maps for mobile robot localization: An evaluation study on cold benchmarking database," pp. 542–547, 2010.

[10] W. Maddern, M. Milford, and G. Wyeth, "Capping computation time and storage requirements for appearance-based localization with CAT-SLAM," pp. 822–827, 2012.

[11] C. Valgren, T. Duckett, and A. Lilienthal, "Incremental spectral clustering and its application to topological mapping," pp. 4283–4288, 2007.

[12] A. Stimec, M. Jogan, and A. Leonardis, "Unsupervised learning of a hierarchy of topological maps using omnidirectional images," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, pp. 639–665, 2007.

[13] 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," *Journal of Sensors*, vol. 2016, pp. 1–21, 2016.

[14] R. González and R. Woods, *Digital image processing*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2008.

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

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients fot human detection." 2005.

[17] M. Hofmeister, M. Liebsch, and A. Zell, "Visual self-localization for small mobile robots with weighted gradient orientation histograms," 2009.

[18] L. Payá, F. Amorós, L. Fernández, and O. Reinoso, "Performance of global-appearance descriptors in map building and localization using omnidirectional vision," *Sensors*, vol. 14, no. 2, pp. 3033–3064, 2014.

[19] A. Oliva and A. Torralba, "Building the gist of ascene: the role of global image features in recognition." 2006.

[20] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *Robotics, IEEE Transactions on*, vol. 25, no. 4, pp. 861–873, 2009.

[21] C.-K. Chang, C. Siagian, and L. Itti, "Mobile robot vision navigation and localization using gist and saliency," pp. 4147–4154, 2010.

[22] A. Murillo, G. Singh, J. Kosecka, and J. Guerrero., "Localization in urban environments using a panoramic gist descriptor," *IEEE Transactions on Robotics*, vol. 29, no. 1, pp. 146–160, 2013.

[23] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," pp. 849–856, 2001.

[25] ARVC. Automation, Robotics and Computer Vision Research Group. Miguel Hernández University. Spain. Quorum 5 set of images. [Online]. Available: http://arvc.umh.es/db/images/quorumv/