






The Role of Global Appearance of Omnidirectional Images in Relative Distance and Orientation Retrieval

Vicente Román ^{*}, Luis Payá , Adrián Peidró , Mónica Ballesta  and Oscar Reinoso 

Department of Systems Engineering and Automation, Miguel Hernández University, 03202 Alicante, Spain; lpaya@umh.es (L.P.); apeidro@umh.es (A.P.); m.ballesta@umh.es (M.B.); o.reinoso@umh.es (O.R.)

^{*} Correspondence: v.roman@umh.es; Tel.: +34-96-665-8859

Abstract: Over the last few years, mobile robotics has experienced a great development thanks to the wide variety of problems that can be solved with this technology. An autonomous mobile robot must be able to operate in a priori unknown environments, planning its trajectory and navigating to the required target points. With this aim, it is crucial solving the mapping and localization problems with accuracy and acceptable computational cost. The use of omnidirectional vision systems has emerged as a robust choice thanks to the big quantity of information they can extract from the environment. The images must be processed to obtain relevant information that permits solving robustly the mapping and localization problems. The classical frameworks to address this problem are based on the extraction, description and tracking of local features or landmarks. However, more recently, a new family of methods has emerged as a robust alternative in mobile robotics. It consists of describing each image as a whole, what leads to conceptually simpler algorithms. While methods based on local features have been extensively studied and compared in the literature, those based on global appearance still merit a deep study to uncover their performance. In this work, a comparative evaluation of six global-appearance description techniques in localization tasks is carried out, both in terms of accuracy and computational cost. Some sets of images captured in a real environment are used with this aim, including some typical phenomena such as changes in lighting conditions, visual aliasing, partial occlusions and noise.

Keywords: omnidirectional imaging; global appearance description; localization; image retrieval; relative orientation; fourier signature; histogram of oriented gradients; gist



Citation: Román, V.; Payá, L.; Peidró, A.; Ballesta, M.; Reinoso, O. The Role of Global Appearance of Omnidirectional Images in Relative Distance and Orientation Retrieval. *Sensors* **2021**, *21*, 3327. <https://doi.org/10.3390/s21103327>

Academic Editor: Radu Danescu

Received: 5 April 2021
Accepted: 4 May 2021
Published: 11 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Nowadays, the presence of mobile robots has increased substantially in many areas, such as industry, households, transportation and education. As their abilities in perception and computation have increased, they have become an efficient tool to perform a wide range of tasks and they are expected to play a crucial role in the development of some activities. In this context, map building and localization are two of the main abilities a robot must develop to be really autonomous. Finding a solution to both problems, balancing accuracy, efficiency and robustness, is very important so that a robot can navigate autonomously and safely through real working environments [1].

In the field of perception, vision sensors have become a widespread tool to get information from the environment [2] due to several factors: the big amount of information they can capture with a relatively low cost; the availability of the data they provide (unlike GPS, whose signal may not be available temporarily, indoors or in narrow outdoor areas); the variety of configurations that they permit, from single-view cameras to binocular or trinocular systems; and the possibility of carrying out other high-level tasks such as people detection. Among the available configurations, catadioptric vision systems stand out thanks to their wide field of view, up to 360 deg around the camera axis [3]. The information captured with these systems can be projected onto varied surfaces, what permits different mathematical approaches depending on the type of task to solve [4]. Omnidirectional

images are particularly effective comparing to conventional images due to the fact that they capture a global context of the environment. Therefore, with this kind of information, global features constitute an effective alternative, compared to local features, to many tasks, such as, for example, the reconstruction of complex indoor environments. In this regard, Sun et al. [5] and Pintone et al. [6] make use of deep learning approaches [7–9] to panoramic image analysis, with the objective of understanding the layout of indoor environments.

Solving the mapping and localization problems using only visual information is challenging. Images are highly dimensional data and they usually contain much redundant information. This information tends to change not only when the robot moves but also under some other usual circumstances such as changes in the external lighting conditions, noise during the acquisition of the image and occlusions due to the presence of, e.g., people in the environment. In addition, when a robot has to operate in indoor environments, it has to cope with the phenomenon of *visual aliasing*, which means that the visual information captured from very different positions may be very similar. Taking these facts into account, to build a functional visual model of the environment and to estimate the pose (position and orientation) of the robot within this model with robustness, it is necessary to find an alternative codification which is more efficient and robust against such phenomena.

Two main frameworks can be found in the literature to extract this information based either on local or on global appearance. The first family of methods consists in detecting some outstanding landmarks or regions and describing them using any algorithm that provides some invariance against transformations, such as SIFT [10], SURF [11], BRIEF [12], BRISK [13], ORB [14], FREAK [15] and LDB [16]. The second family consists of working with each scene as a whole, trying to build a unique descriptor per image that collects information on its global structure, using some approaches such as Principal Components Analysis [17], discrete Fourier transform [18], banks of Gabor filters [19], color histograms [20,21], directly subsampled versions of the original image [22] or Radon transform [23].

Traditionally, researchers have focused on the use of local appearance methods, and it can be considered a mature technology to solve the mapping and localization problems. Many approaches are proposed in the literature based on these descriptors [24–28]. Typically, they require the implementation of detection, description and tracking algorithms which tend to be relatively complex and computationally expensive. While they are often designed to be invariant against some movements of the robot, their behavior can deteriorate when other usual phenomena are present, such as changes in lighting conditions, occlusions, noise or visual aliasing. Some comparative analyses of this kind of descriptor can be found in [29,30]. Thanks to these comparatives, an optimal description method can be chosen and tuned depending on the environment and application.

Global-appearance approaches have been applied to these areas more scarcely. Since each image is described through a unique descriptor, they usually lead to models of the environment that can be handled intuitively by a human operator. The localization process is more straightforward, based on the pairwise comparison between descriptors. Some authors have made use of such approaches in the field of mobile robots, such as [31–36]. These techniques may be useful in unstructured environments where it is difficult to extract robust landmarks. As a drawback, they have been used typically to build topological models [37,38], since no metric information can be extracted from pure global appearance (unless additional sensory information is added).

In [39], a comparative evaluation of the performance of global-appearance methods in mapping tasks was carried out. However, we have not found any work in the literature that makes a deep and systematic study of the role of global appearance in localization tasks. Therefore, the objective of this paper is two-fold. On the one hand, we have chosen six widespread and accepted families of visual description methods, and we have adapted them to be used efficiently with omnidirectional visual information, in such a way that the resulting descriptors contain useful information to retrieve relative distance and orientation efficiently. To this aim, some algorithms have been implemented to estimate the relative

position and orientation from these descriptors using purely visual information. On the other hand, we carry out a comparative evaluation of these descriptors in localization tasks and study their behavior against changes in the robot pose and other visual changes in the environment. Their relative performance has been tested and the influence of the most relevant parameters is assessed, completing the work presented in [39].

The remainder of the paper is structured as follows. Section 2 presents a state-of-the-art of global appearance description approaches and outlines the implementation of the three methods included in the evaluation. After that, in Section 3 the framework used to estimate the position and the orientation of the robot is detailed. Then, Section 4 presents the experimental setup and the set of images used in the experiments. The paper finishes with the results of the experiments, discussed in Section 5, and the conclusions and future lines of research in Section 6.

2. Global Appearance Descriptors

The objective of this section is two-fold. On the one hand, a state-of-the-art of global appearance descriptor is developed. On the other hand, a brief mathematical description of the methods included in the comparative analysis is made. Six families of global appearance methods have been chosen to be analyzed: methods based on the discrete Fourier transform (Section 2.1), on gradient orientation (Section 2.2), on the use of banks of Gabor filters (Section 2.3), on Speeded-Up Robust Features (SURF) description method (Section 2.4), on Binary Robust Independent Elementary Features (BRIEF) (Section 2.5) and on Radon transform (Section 2.6). A complete description of the methods can be found in [39–41]. However, for the sake of clarity, we have included an outline in this section.

We consider the movement of the robot is contained in the ground plane, and it captures images using an omnidirectional vision system mounted on its top. This system consists of a camera pointing towards a hyperbolic mirror, with their axes aligned and in vertical position. The complete experimental setup is presented in Section 4.

2.1. Descriptors Based on the Discrete Fourier Transform

The discrete Fourier transform (DFT) has been used by many researchers to extract the most relevant information from scenes. For example, Oliva and Torralba [19] propose using a windowed 2D Fourier transform, that permits defining some circular windows to select spatial information around some specific pixels in the scene. Ishiguro and Tsuji [42] propose an alternative approach, named Fourier Signature (FS), which is designed to be used on panoramic images. Menegatti et al. showed the robustness of this representation to build a model of an environment and to estimate the position of a vehicle using a Monte Carlo approach [18,31], in a relatively small environments and controlled conditions. Stürzl et al. [43] propose a visual homing algorithm based on the Fourier Signature, but the panoramic scene is previously reduced to a unidimensional array. Horst and Möller use it in visual place recognition [44].

The Fourier Signature (FS) permits obtaining a descriptor which is invariant against rotations of the robot in the ground plane when using panoramic images. For this reason, this is the DFT-based representation we have chosen in this comparative evaluation. The description process starts from a panoramic scene $f(x, y) \in \mathbb{R}^{N_1 \times N_2}$. Initially, the image can be subsampled to obtain a lower number of rows $k_1 < N_1$ ($k_1 = 1$ in [43]). The FS of the resulting scene $f(x, y) \in \mathbb{R}^{k_1 \times N_2}$ is the matrix $\mathbf{F}(u, y) \in \mathbb{C}^{k_1 \times N_2}$ obtained after calculating the unidimensional DFT of each row of the image. In the frequency domain, the main information is concentrated in the low frequency components, and the high frequency components tend to be more contaminated by the possible presence of noise in the original image. Taking this fact into account, by retaining the k_2 first columns and discarding the remainder, a compression effect is achieved. The new complex matrix, with k_1 rows and k_2 columns, can be expressed as a magnitudes matrix $\mathbf{A}(u, y) = \|\mathbf{F}(u, y)\|$ and an arguments matrix $\Phi(u, y)$.

Based on the shift Theorem of the unidimensional DFT, when two panoramic images have been captured from the same point on the floor, but having the robot different orientations around the vertical axis, both images present the same magnitudes matrix, and the arguments matrices can be used to estimate the relative orientation of the robot. Thanks to this property, the matrix $\mathbf{A}(u, y) = \|\mathbf{F}(u, y)\|$ can be considered as a visual descriptor of the robot position (as it is rotationally invariant), the matrix $\Phi(u, y)$ can be considered as a descriptor of the robot orientation (as it permits estimating this orientation), and the estimation of the position and the orientation can be addressed independently and sequentially.

To sum up, the position descriptor is the matrix $\mathbf{A}(u, y) \in \mathbb{R}^{k_1 \times k_2}$ and the orientation descriptor is the matrix $\Phi(u, y) \in \mathbb{R}^{k_3 \times k_4}$. In the experiments, different sizes will be considered, to test separately the influence these parameters have on the accuracy and computational cost of the localization process.

2.2. Descriptors Based on Histograms of Oriented Gradients

The Histograms of Oriented Gradients (HOG) are local descriptors that have been used typically in computer vision and image processing to solve object detection tasks. HOG was initially described by Dalal and Triggs [45], who used it to detect persons in sequences of images. Afterwards, some researchers presented an improved version both in detection and computational cost [46]. Hofmeister et al. [47] made use of HOG to solve the localization of small mobile robots from low resolution images, in visually simple environments and when the orientation of the robot is similar to the orientation it had when the corresponding map image was captured. In [48], the same authors present a comparative of HOG with other appearance descriptors, applied to the localization of small robots in reduced environments, with similar results. Aslan et al. study the ability of HOG to handle occlusion in human tracking [49]. In addition, Neumann et al. use HOG, among other descriptors, for image-based vehicle detection and localization in an autonomous car [50].

Originally, HOG is built to describe local areas of a scene. We redefine it as a global appearance descriptor, using an exhaustive set of cells that covers the whole image and permits describing the global appearance. The version of HOG included in the comparative evaluation is presented in [51], where a global version of HOG is used to carry out map building and Monte Carlo localization in a large environment. When used to describe panoramic scenes, it presents rotational invariance and it also permits estimating the orientation of the robot.

In brief, from the initial panoramic image, a position and an orientation descriptor are obtained using the HOG philosophy. From the initial panoramic image $f(x, y) \in \mathbb{R}^{N_1 \times N_2}$ the magnitude and the orientation of the gradient are obtained and stored in the matrices $\mathbf{M}(x, y)$ and $\Theta(x, y)$, respectively. From now on, some sets of cells are defined upon the matrix $\Theta(x, y)$ to build the two descriptors. On the one hand, to build the position descriptor, a set of k_5 horizontal cells, whose width is equal to N_2 pixels, without overlapping, and covering the whole image are defined. For each cell, an orientation histogram with b_1 bins is compiled. During this process, each pixel in $\Theta(x, y)$ is weighted with the magnitude of the corresponding pixel in $\mathbf{M}(x, y)$. At the end of the process, the set of histograms are appended to compose the position descriptor $\vec{h}_1 \in \mathbb{R}^{k_5 \cdot b_1 \times 1}$. On the other hand, the orientation descriptor is built using the same steps, but considering a set of overlapped vertical cells, with a height equal to N_1 pixels, width equal to l_1 and distance between two consecutive cells equal to d_1 . The number of vertical cells is $k_6 = N_2/d_1$. After compiling a gradient orientation histogram for each cell, with b_2 bins and appending them, the result is the orientation descriptor $\vec{h}_2 \in \mathbb{R}^{k_6 \cdot b_2 \times 1}$.

The descriptor \vec{h}_1 is invariant against rotations of the robot in the ground plane so it can be considered as a visual descriptor of the robot position, and the information contained in \vec{h}_2 permits estimating the orientation of the robot with respect to a reference image.

2.3. Descriptors Based on Gist

The descriptors based on *gist* try to imitate the ability of the human perception system to recognize immediately a scene through the identification of specific regions stand out with respect to their neighborhood. This concept was introduced by Oliva and Torralba [52,53] with the idea of creating a low dimensional global image descriptor. More recent works make use of the concept of *prominence* together with *gist*. Siagian et al. [54] try to establish synergies between both concepts in a unique descriptor whose computational cost is relatively reduced. While these descriptors have been used thoroughly in classification tasks, the experience in mobile robotics localization is more sparse. Some related applications can be found in [55], where a localization and navigation system based on the *gist* and *prominence* concepts is presented; in [56], where *gist* descriptors, calculated over specific portions of a set of panoramic images, are used to solve a localization problem in urban areas; and in [57], where descriptors based on *gist* and dimensionally reduced by means of Principal Components Analysis are used to solve the loop closure problem in Simultaneous Localization and Mapping. In addition, Su et al. use *gist* in a localization framework to match keyframes, in combination with local descriptors to improve localization accuracy [58].

The description method we have included in this comparative analysis is based on the works of Siagian et al. [54] and is deeply described in [51]. It is built from orientation information, obtained by means of a bank of Gabor filters with different orientation, in some levels of resolution. First, two versions of the original panoramic image are considered: the original one and a new lower resolution version after applying a Gaussian low-pass filter and subsampling to a new size $0.5 \cdot N_1 \times 0.5 \cdot N_2$. After that, both images are filtered with a bank of m_1 Gabor filters whose orientations are evenly distributed between 0 and 180 deg. Finally, to reduce the amount of information, the pixels in each resulting image are grouped into blocks, by calculating the average intensity of all the pixels contained in a block. The block division is chosen in an identical fashion than in the case of HOG. First, a set of k_7 horizontal blocks is defined to obtain the position descriptor $\vec{g}_1 \in \mathbb{R}^{2 \cdot k_7 \cdot m_1 \times 1}$, which is invariant against rotations of the robot in the ground plane. Second, a set of k_8 vertical blocks with overlapping is defined to obtain the orientation descriptor $\vec{g}_2 \in \mathbb{R}^{2 \cdot k_8 \cdot m_2 \times 1}$.

2.4. Descriptors Based on Wi-SURF

SURF [11] has been considered one of the most important local descriptors and it has been used in countless works as in [59] or [32] where it is used to solve localization indoors. The present study is focused on the performance of global appearance descriptors. For this reason, we propose an adaptation which is based on the work [60], which extracts a unique, global appearance descriptor per image, using the SURF philosophy. Throughout the paper, we will refer to this descriptor as Whole Image SURF (Wi-SURF).

Wi-SURF has been used in previous works for topometric localization [61] or for place recognition [40]. These works propose to obtain a unique vector $d \in \mathbb{R}^{64}$ that contains gradient information of the entire image. Therefore, such a descriptor can be useful for place recognition, but does not contain enough information to estimate relative orientation. For this reason, we propose dividing the panoramic image into a set of evenly distributed square windows, with some overlapping between them. In each window, a SURF descriptor $d \in \mathbb{R}^{64}$ is calculated and all the descriptors are concatenated, which leads to a global-appearance descriptor. This approach will enable us to solve not only the localization but also to estimate the relative orientation of the robot, as detailed in Section 3.4. The square windows are evenly distributed following the next parameters: k_9 is the number of horizontal cells in which the panoramic image is split and sp_1 the horizontal space between consecutive windows. The number of windows per cell will depend on the images' width (512 columns in our experiments) so a total of $w_1 = \frac{512}{sp_1}$ windows per cell are calculated. The width of the square window is equal to the height of the horizontal cell. After all, the size of the descriptor is $\vec{w}s \in \mathbb{R}^{k_9 \cdot w_1 \cdot 64 \times 1}$. This final descriptor will be used to estimate both position and orientation.

2.5. Descriptors Based on BRIEF-Gist

BRIEF-gist is a global appearance descriptor based on the local descriptor Binary Robust Independent Elementary Features (BRIEF). BRIEF was presented in [12] and used for different mobile robot applications [62,63]. Based on this local descriptor, a global appearance descriptor is presented in [64]. This approach is known as BRIEF-gist and it has been used for place recognition and loop closure detection in [40]. In the present work, we adapt this descriptor to be used with panoramic images in such a way that it permits calculating both relative distance and orientation in a localization task.

To implement the BRIEF-gist descriptor, the image is divided into $k_{10} \times w_2$ windows equally sized. Then, using the BRIEF description methodology, a set of ordered pairs of pixels is defined in each window, and the intensity of the second pixel of each pair is compared to the first one. If the difference is positive a 1 is added to the global descriptor, and a 0 if the difference is negative. As a result, a boolean vector is obtained. After this process, the resulting BRIEF-gist descriptor is $\vec{b}g \in \mathbb{R}^{k_{10} \cdot w_2 \times 1}$. This final descriptor is used to estimate both position and orientation.

2.6. Descriptors Based on Radon Transform

The Radon transform was proposed in [65]. Initially, it was used in different computer vision applications as a geometric shape descriptor, as in [66,67]. More recently, the Radon transform (RT) has been adapted to describe globally omnidirectional images and its performance was tested in [41], where descriptors based on the RT were used to solve the image retrieval problem, and in [23], where these descriptors were used to estimate relative altitude from images. The main advantage of this descriptor is that it can be calculated with raw omnidirectional images, as captured by the vision system (with no panoramic transformation).

Mathematically, the Radon transform consists of describing a function in terms of the projections of its linear integrals.

After applying the Radon transform, the image is transformed into a function $r_{im}(\Phi, d)$, which is obtained after integrating the original function through several groups of parallel lines with distance to the origin d and different orientation Φ . The size of the new descriptor is $r_{im} \in \mathbb{R}^{M_x \times M_y}$, M_x is the number of orientations where $\Phi = \{\Phi_1, \Phi_2, \dots, \Phi_{M_x}\}$ and M_y is the number of parallel lines.

When the Radon transform is applied to omnidirectional images, it is specially interesting its symmetry and the fact that the descriptor is horizontally shifted when the robot rotates [68], which allows us to obtain global appearance descriptors that can be used to estimate position and relative orientation. This property can be seen in Figure 1, where four omnidirectional images are shown; three of them have been taken from the same position but with different orientation and the other one has been taken from a different position. The figure clearly shows the effect of the orientation in the Radon transform and how different the result is if the image is from another room. If the robot rotates $(\Delta\theta)$ degrees, the new descriptor presents the same information as the original one, but it has been shifted s columns, $s = (\Delta\theta) \cdot (M_x)/360$. Thanks to this property, descriptors based on Radon transform contain position and orientation information of the robot.

To sum up, after applying the Radon transform to an omnidirectional image with size $N_x \times N_x$, a matrix $r \in \mathbb{R}^{\frac{360}{p_1} \times 0.5 \cdot N_x}$ is obtained. p_1 is the angle (deg.) between consecutive sets of lines along which the linear integrals are calculated. In the experiments, these matrices can be used in different ways in order to obtain proper uni-dimensional descriptors. Two different methods and different sizes will be considered to test the robustness of the descriptor in pose estimation. These methods and parameters are described in Section 3.

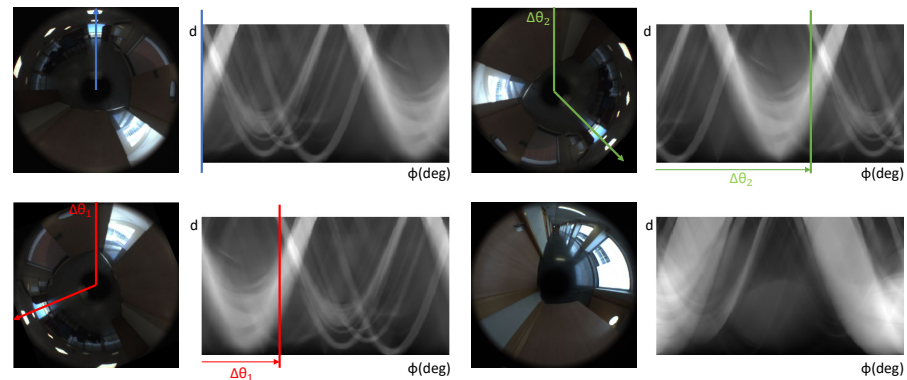


Figure 1. Shift property in the Radon transform.

3. Solving the Absolute Localization Problem

In this work, we assume a visual model of the environment is previously available. To build this model, the robot has gone through the initially unknown environment (either in a tele-operated way or using any exploration algorithm [69,70]) and has captured a set of omnidirectional images from n points of view, defined by the poses $\vec{p}_j = (x_j, y_j, \theta_j)$, $j = 1, \dots, n$, to cover the whole environment to map. The model \mathcal{M} is composed of the visual descriptors and the pose of the robot, stored for each capture position: $\mathcal{M} = \{(\mathcal{D}_1, \vec{p}_1), (\mathcal{D}_2, \vec{p}_2), \dots, (\mathcal{D}_n, \vec{p}_n)\}$ where, in general, the description of each image consists of a position and an orientation descriptor $\mathcal{D}_j = \{\vec{d}_{1j}, \vec{d}_{2j}\}$ (in the case of *Wi-SURF* and *BRIEF-gist* the same vector is used as position and orientation descriptor, so $\mathcal{D}_j = \{\vec{d}_{1j}\}$). The map building process using global appearance methods and omnidirectional imaging is thoroughly described in [39].

Once the model is built, the localization problem consists of estimating the pose of the robot. The problem is approached here as an absolute localization problem, i.e., no information on the previous position of the robot is considered, and only visual information is used. The robot captures a new image at time instant t , from an unknown pose (f_t , *test image*). Then, the descriptor of this image \mathcal{D}_t is computed and compared with the set of descriptors stored in the model. From this comparison, the position and orientation of the robot at time instant t are estimated. The next subsections detail these processes depending on the description method used.

3.1. Descriptors Based on the Discrete Fourier Transform

When a test image arrives, \mathbf{A}_t and Φ_t are calculated. Since the position descriptor is invariant against rotations of the robot in the ground plane, first, \mathbf{A}_t is used to estimate the position of the robot, by comparing it with the descriptors \mathbf{A}_j , $j = 1, \dots, n$ and retaining the k -nearest neighbors. The position of the nearest neighbor (x_i, y_i) (i is the index of the nearest neighbor) can be considered as an estimation of the position of the robot at time instant t . Once the position of the robot has been estimated, the arguments matrix of the *test image*, Φ_t , and the arguments matrix of the nearest neighbor, Φ_i , are used to estimate the orientation of the robot, using the shift theorem of the DFT. The objective is to estimate the relative orientation θ_{ti} of the robot at time instant t with respect to the orientation the robot had when capturing the nearest neighbor, $\theta_{ti} = \theta_t - \theta_i$. The next steps are as follows:

1. A set of artificial rotations is applied to the *test image*. The shift theorem of the unidimensional DFT can be used to generate the argument matrices of the test image rotated siblings. The step between consecutive rotations is $\Delta\phi$. This is equivalent to making a shift of the columns of the panoramic image with a magnitude of d pixels, where $\Delta\phi = d \cdot 2\pi/N_2$. In the experiments, we consider $d = \{1, 2, \dots, N_2 - 1\}$. This means that the angular step between consecutive artificial rotations is $\Delta\phi = 2\pi/N_2$. This is the resolution of the method.

- After this process, a set of $n_{rot} = 2\pi/\Delta\phi$ arguments matrices are available at time instant t .

$$\{\Phi_0, \Phi_1, \dots, \Phi_{n_{rot}}\}_t = \{\Phi_\alpha\}_t, \alpha = 0, \dots, n_{rot} \quad (1)$$

- The Hadamard product of the matrix Φ_t and every matrix Φ_α is calculated. The sum of the components of each resulting matrix is obtained, and the result is an array of data:

$$\{m_0, m_1, \dots, m_{n_{rot}}\}_t = \{m_\alpha\}_t, \alpha = 0, \dots, n_{rot} \quad (2)$$

- The estimated relative rotation is the α value whose coefficient m_α presents the maximum value.

$$\alpha = \arg \max_\alpha \{m_\alpha\} \quad (3)$$

$$\theta_{ti} = \frac{2\pi\alpha}{n_{rot}} \quad (4)$$

where θ_{ti} is the relative orientation between the image im_t and the nearest neighbor of the map, im_i . This way, the absolute orientation of the robot at time instant t can be calculated as:

$$\theta_t = \theta_i + \theta_{ti} \quad (5)$$

In this equation, θ_i is the orientation that the robot had when the map image im_i was captured, with respect to the global reference system.

In the experiments, the parameters of the Fourier Signature to optimize are the size of the module matrix (k_1 and k_2) and the size of the arguments matrix (k_3 and k_4) to reach a balance between the accuracy in the estimation of the position and orientation and the computational cost of the algorithms.

3.2. Descriptors Based on Histograms of Oriented Gradients

Once the test image im_t has been captured, the descriptors \vec{h}_{1t} and \vec{h}_{2t} are calculated. First, the k -nearest neighbors to \vec{h}_{1t} among the set of descriptors \vec{h}_{1j} , $j = 1, \dots, n$ are calculated and extracted. The position (x_i, y_i) of the nearest neighbor i is an estimation of the position of the robot at time instant t .

Later, the orientation is calculated by comparing the vector \vec{h}_{2t} with the vector \vec{h}_{2i} . With this aim, a set of artificial rotations is calculated using the vector \vec{h}_{2t} and later, the scalar product between the resulting vector after each rotation and the vector \vec{h}_{2i} is calculated. To simulate a rotation of the vector \vec{h}_{2t} , the circular shift must be a multiple of b_2 positions (b_2 is the number of bins per histogram). A shift of b_2 positions equals a rotation of the robot $\Delta\phi = 2\pi d_1/N_2$ radians (this is the angular resolution of the method), where d_1 is the distance between two consecutive vertical cells.

Finally, the estimated relative orientation θ_{ti} of the robot is the angle that corresponds to the rotated version of the vector \vec{h}_{2t} which presents a higher scalar product with \vec{h}_{2i} .

3.3. Descriptors Based on Gist

The processes to estimate the position and orientation are identical to those presented in the case of HOG. Once captured the test image im_t , the descriptors \vec{g}_{1t} and \vec{g}_{2t} are calculated. First, \vec{g}_{1t} is compared to \vec{g}_{1j} , $j = 1, \dots, n$ and the k -nearest neighbors are calculated. From them, the position $(x, y)_i$ of the nearest neighbor i is considered an estimation of the position of the robot at time instant t . After that, the orientation is calculated by comparing the vector \vec{g}_{2t} with the vector \vec{g}_{2i} . With this aim, successive artificial rotations are calculated, using the vector \vec{g}_{2t} and later, the scalar product between

each rotated version and the vector \vec{g}_{2i} is obtained. To make an artificial rotation of the vector \vec{h}_{2i} , the magnitude of the circular shift must be a multiple of m_2 (m_2 is the number of components of each vertical block). Every shift equals a rotation of $\Delta\phi = 2\pi d_2/N_2$ radians (this is the angular resolution of the method), where d_2 is the distance between two consecutive vertical blocks in the descriptor.

The resulting orientation θ_{ti} is the angle that corresponds to the rotated version of \vec{g}_{2i} that presents the highest scalar product with \vec{g}_{2i} .

3.4. Descriptors Based on Wi-SURF

Once the test image im_t is taken, the descriptor \vec{w}_{s_t} is obtained. First, this descriptor is compared with the descriptors $\vec{w}_{s_j}, j = 1, \dots, n$, to calculate the relative orientation between the test descriptor and the descriptors in the model. To estimate the relative orientation, some artificial rotations are added to \vec{w}_{s_t} and the distance between the resulting descriptor after each rotation and \vec{w}_{s_j} is calculated. To simulate an artificial rotation of \vec{w}_{s_t} , a circular shift is applied, which must be a multiple of 64 positions (the SURF descriptor of each window contains 64 components) and w_1 (number of windows). The 64-position shift of the descriptor equals to a rotation of the robot $\Delta\phi = 2 \cdot \pi \cdot sp_1/N_2$ radians (and therefore, this is the angular resolution of the method). Once the relative orientation between the test descriptor and each of the descriptors in the model has been calculated, each descriptor \vec{w}_{s_j} is shifted in such a way that the resulting descriptor has the same orientation as \vec{w}_{s_t} .

Once all the descriptors are supposed to be in the same orientation, the k -nearest neighbors to \vec{w}_{s_t} are calculated among the set of descriptors in the model (once they are equally oriented with respect to \vec{w}_{s_t}). The position (x_i, y_i) of the nearest neighbor i is an estimation of the position of the robot at time t . The orientation between them has been calculated previously and the corresponding angle θ_{ti} is the relative orientation estimated between the test vector and the vector evaluated from \vec{w}_{s_j} .

3.5. Descriptors Based on BRIEF-Gist

Firstly, the relative orientation between images is estimated. The descriptor \vec{b}_{g_t} is calculated from the test image im_t , and the relative orientation between it and each of the descriptors $\vec{b}_{g_j}, j = 1, \dots, n$ is estimated. To estimate it, successive artificial rotations are applied to \vec{b}_{g_t} , the scalar product between the resulting vector after each rotation and \vec{b}_{g_j} is calculated and the minimum is retained. To simulate an artificial rotation of \vec{b}_{g_t} , the circular shift must be a multiple of w_2 (number of windows in each cell). As explained in Section 2.5, to calculate this descriptor the image is divided into $k_{12} \times w_2$ windows, so the angular resolution of the method is determined by the number of windows w_2 . Every w_2 shift is equal to a rotation of the robot $\Delta\phi = 2 \cdot \pi/w_2$ radians.

After estimating the relative orientation, each descriptor in the model \vec{b}_{g_j} is rotated such that the resulting descriptor has the same orientation than \vec{b}_{g_t} . Then the k -nearest neighbors to \vec{b}_{g_t} are calculated among the set of rotated descriptors in the model. The position (x_i, y_i) of the nearest neighbor i is an estimation of the position of the robot at time t . The relative orientation between them has been calculated previously and the corresponding angle θ_{ti} is the difference of orientation estimated between the test vector and the vector evaluated from \vec{w}_{s_j} .

3.6. Descriptors Based on the Radon Transform

In the present work, we process this descriptor using two different methods to retrieve both position and orientation.

3.6.1. Radon–Fourier Method

After applying the Radon transform, a matrix $r \in \mathbb{R}^{\frac{360}{p_1} \times 0.5 \cdot N_x}$ is obtained. Then, the Fourier Signature of this matrix is calculated. As a result of this second transformation,

a matrix of magnitudes $\mathbf{A}_{RTj} \in \mathbb{R}^{\frac{360}{p_1} \times k_{11}}$ and a matrix of arguments $\Phi_{RTj} \in \mathbb{R}^{\frac{360}{p_1} \times k_{12}}$ are obtained. As in the case of the descriptors based on the DFT, \mathbf{A}_{RTj} is used as position descriptor and Φ_{RTj} is used as an orientation descriptor. k_{11} is the number of columns taken for the position descriptor \mathbf{A}_{RTj} and k_{12} is the number of columns taken for the orientation descriptor Φ_{RTj} . To estimate the position and orientation, we use the same process as in the descriptors based on the discrete Fourier transform, presented in the Section 3.1.

3.6.2. Radon–POC Method

This method uses directly the matrix obtained after applying the Radon transform ($r \in \mathbb{R}^{\frac{360}{p_1} \times 0.5 \cdot N_x}$) as the image descriptor r_{poc_j} . To compare two descriptors, Phase Only Correlation (POC) is used. This operation outputs a correlation coefficient that allows us to estimate the similarity between two matrices and their relative shift.

To sum up, Table 1 shows the parameters whose influence will be studied in the comparative evaluation. After that, Table 2 gives details of the contents of the model when we consider each description method.

Table 1. Parameters whose influence in the localization process is studied.

Descriptor	Parameters
<i>FS</i>	$k_1 \Rightarrow$ number of rows, position descriptor \mathbf{A}_j $k_2 \Rightarrow$ number of columns, position descriptor \mathbf{A}_j $k_3 \Rightarrow$ number of rows, orientation descriptor Φ_j $k_4 \Rightarrow$ number of columns, orientation descriptor Φ_j
<i>HOG</i>	$b_1 \Rightarrow$ number of bins per histogram, position descriptor \vec{h}_{1j} $k_5 \Rightarrow$ number of horizontal cells, position descriptor \vec{h}_{1j} $b_2 \Rightarrow$ number of bins per histogram, orientation descriptor \vec{h}_{2j} $l_1 \Rightarrow$ width of vertical cells, orientation descriptor \vec{h}_{2j} $d_1 \Rightarrow$ distance between vertical cells, orientation descriptor \vec{h}_{2j} $k_6 = \frac{N_2}{d_1} \Rightarrow$ number of vertical cells, orientation descriptor \vec{h}_{2j}
<i>Gist</i>	$m_1 \Rightarrow$ number of orientations (Gabor filters), position descriptor \vec{g}_{1j} $k_7 \Rightarrow$ number of horizontal blocks, position descriptor \vec{g}_{1j} $m_2 \Rightarrow$ number of orientations (Gabor filters), orientation descriptor \vec{g}_{2j} $l_2 \Rightarrow$ width of vertical blocks, orientation descriptor \vec{g}_{2j} $d_2 \Rightarrow$ distance between vertical blocks, orientation descriptor \vec{g}_{2j} $k_8 = \frac{N_2}{d_2} \Rightarrow$ number of vertical blocks, orientation descriptor \vec{g}_{2j}
<i>WS</i>	$w_1 \Rightarrow$ number of windows per cell, descriptor $\vec{w}s_j$ $k_9 \Rightarrow$ number of horizontal blocks, descriptor $\vec{w}s_j$ $sp_1 \Rightarrow$ horizontal space between windows, descriptor $\vec{w}s_j$
<i>BG</i>	$w_2 \Rightarrow$ number of windows per cell, descriptor $\vec{b}g_j$ $k_{10} \Rightarrow$ number of horizontal blocks, descriptor $\vec{b}g_j$
<i>RT</i>	$p_1 \Rightarrow$ degrees between lines where Radon is calculated, matrix r $k_{11} \Rightarrow$ number of columns, position descriptor \mathbf{A}_{RTj} $k_{12} \Rightarrow$ number of columns, orientation descriptor Φ_{RTj} $N_x \Rightarrow$ omnidirectional images' size is $N_x \times N_x$

Table 2. Contents of the map, for localization and orientation estimation, per image included in the model im_j , $j = 1, \dots, n$.

Descriptor	Localization	Orientation
FS	$\mathbf{A}_j \in \mathbb{R}^{k_1 \times k_2}$	$\Phi_j \in \mathbb{R}^{k_3 \times k_4}$
HOG	$\vec{h}_{1j} \in \mathbb{R}^{k_5 \cdot b_1 \times 1}$	$\vec{h}_{2j} \in \mathbb{R}^{k_6 \cdot b_2 \times 1}$
Gist	$\vec{g}_{1j} \in \mathbb{R}^{2 \cdot k_7 \cdot m_1 \times 1}$	$\vec{g}_{2j} \in \mathbb{R}^{k_8 \cdot m_2 \times 1}$
WS	$\vec{w}s_j \in \mathbb{R}^{k_9 \cdot w_1 \cdot 64 \times 1}$	
BG	$\vec{b}g_j \in \mathbb{R}^{k_{10} \cdot w_2 \times 1}$	
RT-F	$\mathbf{A}_{RTj} \in \mathbb{R}^{\frac{360}{p_1} \times k_{11}}$	$\Phi_{RTj} \in \mathbb{R}^{\frac{360}{p_1} \times k_{12}}$
RT-POC	$r \in \mathbb{R}^{\frac{360}{p_1} \times 0.5 \cdot N_x}$	

4. Experimental Setup

This section describes the experimental setup. First, the sets of images used to carry out the experiments are presented. Second, a variety of phenomena (noise and occlusions) to test the robustness of the algorithms are described.

4.1. Sets of Images

All the experiments are carried out with two sets of images captured by ourselves [71]. A catadioptric vision system is used to capture the images. It is composed of an *Imaging Source DFK 21BF04* camera pointing towards an *Eizoh Wide 70* hyperbolic mirror, with their axes aligned. This system captures omnidirectional images which are preprocessed to obtain cylindrical projections (panoramic images) with size $N_1 \times N_2 = 128 \times 512$ pixels.

The first set of images is named the *training set* and it is composed of 872 panoramic images captured on a dense grid of points of 40×40 cm, covering the whole floor of a building of Miguel Hernández University (Spain), including 6 different rooms. The *training set* will be used to build a visual model of the environment. Different grid sizes will be considered along the experiments.

The second set is named the *test set* and it contains 1232 images captured in all the rooms, with different orientations. To capture these images, 77 positions were defined on some half-way points among the grid positions, and 16 images per position were captured, with different robot orientations, to cover the whole circumference. These images were captured in different times of day and with changes in the position of some objects, doors, etc., to reflect the natural variability of the visual information in real working environments. The *test set* will be used during the process of localization and orientation estimation, to test the goodness of each description method and the influence of the main parameters. This environment is very prone to *perceptual aliasing*, which means that two images captured from two positions which are far away may have a similar visual appearance. Global appearance descriptors must cope with this phenomenon as it frequently happens in indoor environments.

Figure 2 shows a bird's eye view of the environment and the capture points of the training images. As an example, Figure 3 shows the library, the capture points of the training (red) and test (green) images and some sample training and test images captured in close points. The effect of changes in lighting conditions and changes of orientation can be appreciated. Other sample space is shown in Figure 4 (corridor). The effect of *visual aliasing* is clearly shown. In addition, the test image 3 shows an example of changes in the environment (open door with respect to the training images).

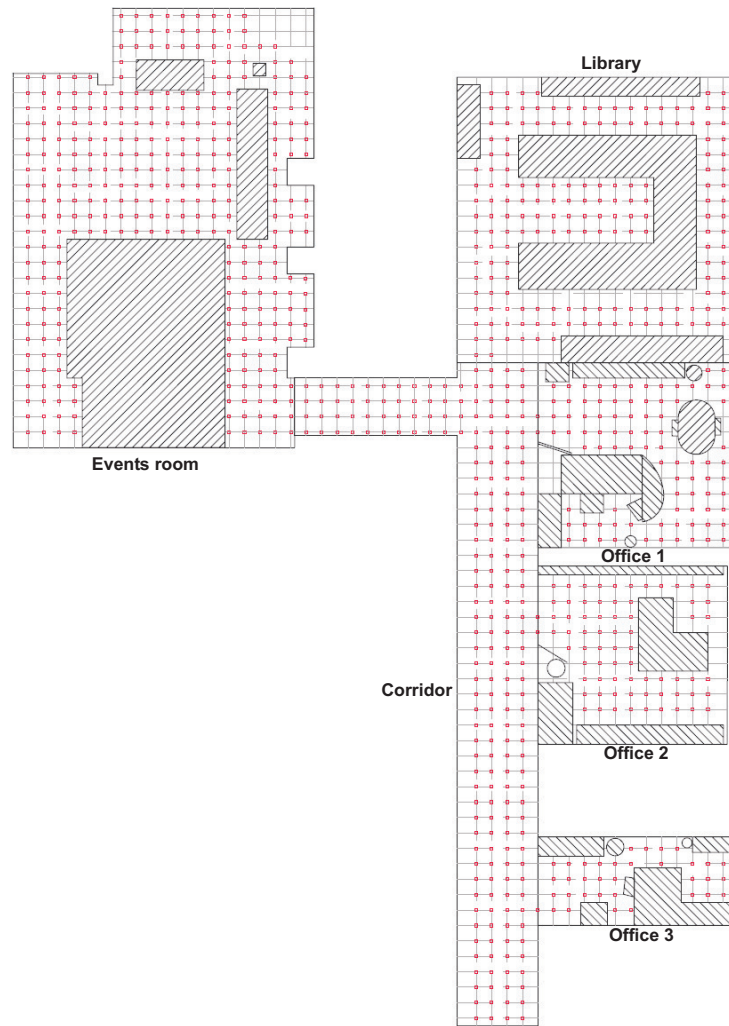


Figure 2. Bird’s eye view of the capture points of the training set of images. The size of the grid is 40×40 cm.

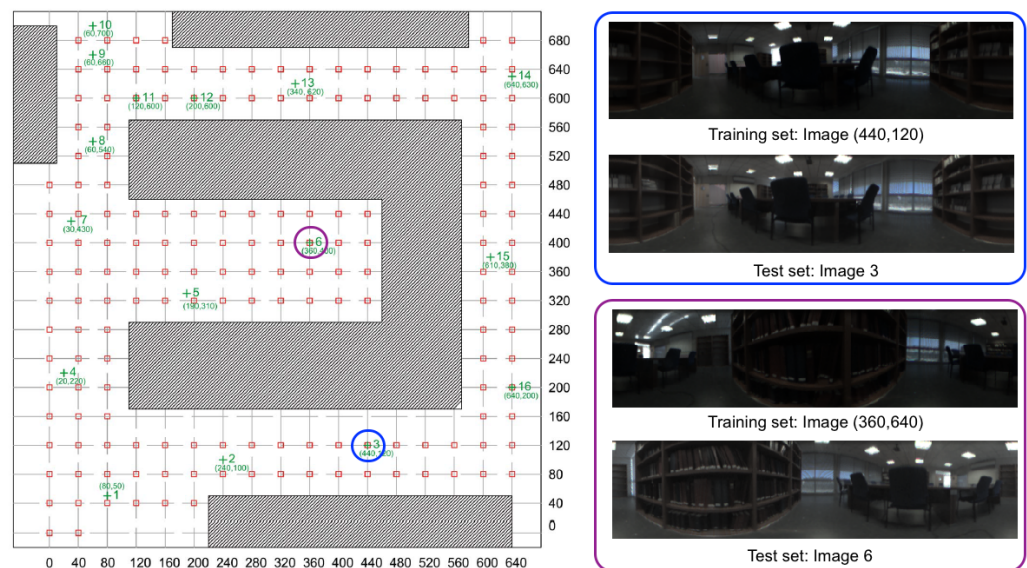


Figure 3. Library. Bird’s eye view of the capture points of the training set of images. The size of the grid is 40×40 cm.

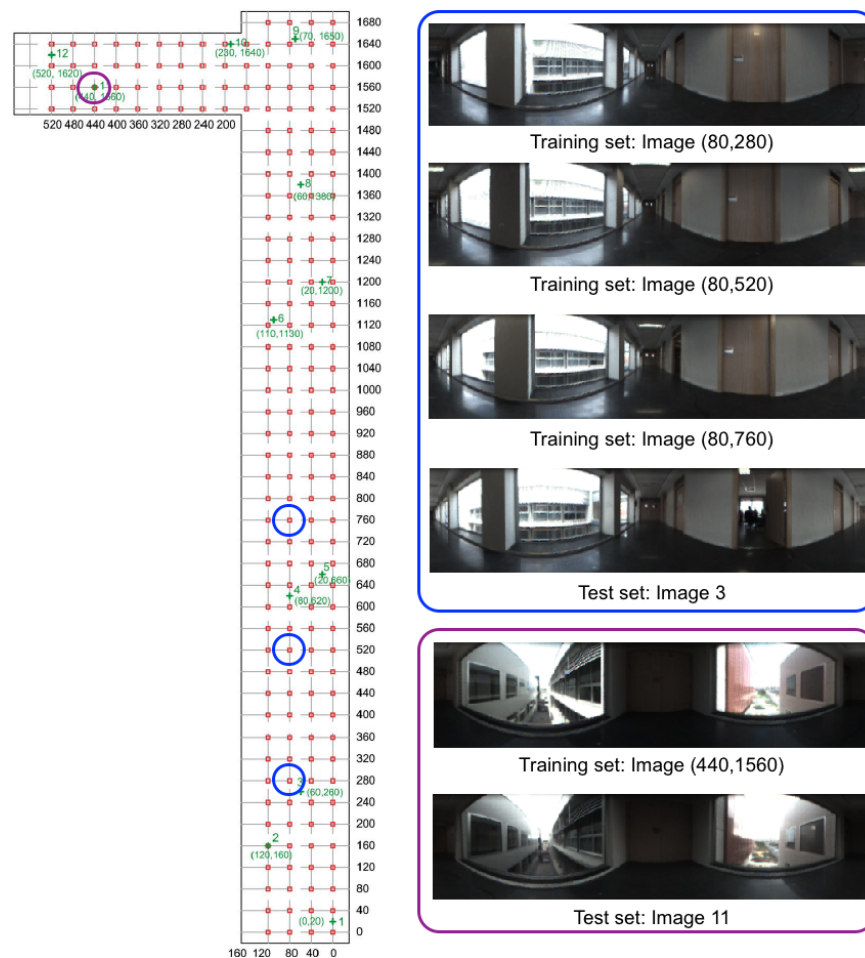


Figure 4. Corridor. Bird's eye view of the capture points of the training set of images. The size of the grid is 40×40 cm.

4.2. Addition of Noise and Occlusions

The test images reflect some of the most habitual undesired effects in real working environment: changes in lighting conditions, in the position and state of some objects and perceptual aliasing. Additionally, two other phenomena are considered in the experiments: noise and occlusions.

First, to test the influence of noise due to the nature of the acquisition system, noise with Gaussian distribution is considered, with null average value and several variance values, to consider different noise levels: $\sigma^2 = \{0, 0.0025, 0.05, 0.01, 0.02, 0.05\}$. Along the rest of the paper, these levels of noise are named noises 0, 1, 2, 3, 4 and 5, respectively. Figure 5a shows a test image with these levels of added noise. In the most extreme case, the visual appearance of the image is seriously altered.

Second, the presence of persons or other robots in the environment may occlude partially and temporarily the visual information. Working with panoramic images constitutes an advantage as far as occlusions are concerned. However, they may hide some relevant features with respect to the visual information stored in the map and put in risk the localization process. To model this effect, several levels of occlusion have been added artificially to the images, considering several vertical bars that produce different levels of occlusion, considering $\{0, 5, 10, 20, 40\}\%$ of the whole image occluded. Along the rest of the paper, these levels are named occlusions 0, 1, 2, 3, 4 and 5, respectively. Figure 5b shows a test image with these levels of added occlusion. In the most extreme case, 40% of the visual information is lost.

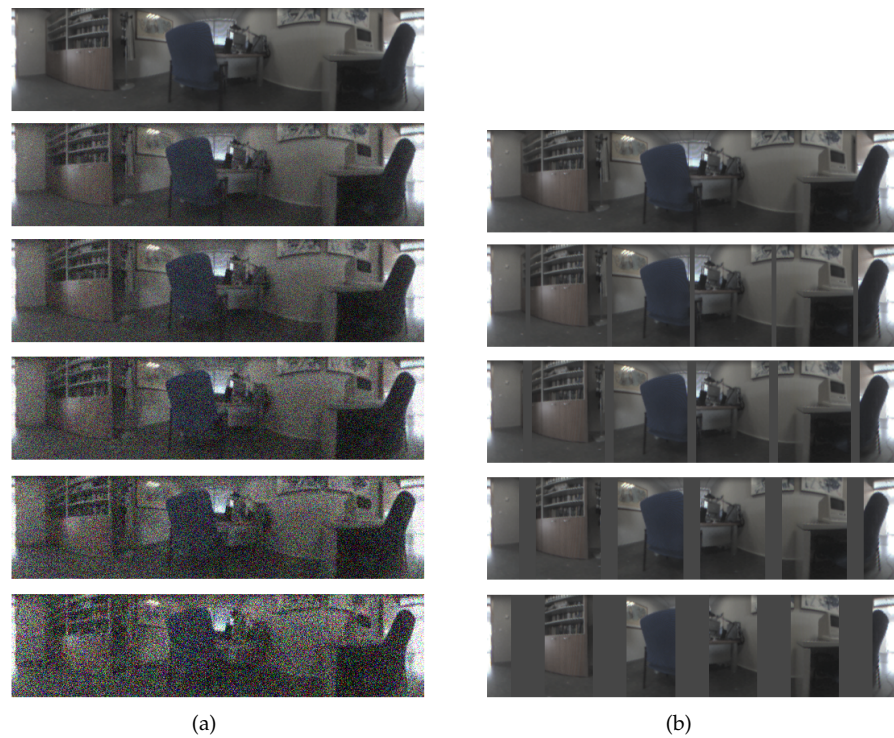


Figure 5. Sample image from the training test with (a) different levels of added Gaussian noise ($\sigma^2 = \{0, 0.0025, 0.05, 0.01, 0.02, 0.05\}$) and (b) sequence of occlusions considered ($\{0, 5, 10, 20, 40\}\%$).

5. Results and Discussion

In this section, an exhaustive bank of experiments is proposed to test the performance of the global appearance descriptors included in the comparative evaluation and the influence of the main parameters in the accuracy and computational cost of the localization process. The experiments have been structured in four subsections. First, in Section 5.1, the ability of each descriptor to find the nearest neighbor of the model, in ideal conditions (considering neither noise nor occlusions) is tested. After that, the problem of position estimation is solved, including also the study of performance with these effects (Section 5.2). Third, in Section 5.3, the problem of orientation estimation is considered. Finally, Section 5.4 studies the relative performance of the descriptors with a trajectory-like dataset.

5.1. Image Retrieval Problem

During the localization process, the first step consists of comparing the localization descriptor of the test image with all the localization descriptors in the map and obtaining the k-nearest neighbors. Taking this fact into account, in this section we evaluate the ability of each description method to calculate correctly the first nearest neighbor (i.e., to identify correctly the position of the model which is geometrically the nearest one to the test position). It is known as the *image retrieval problem*.

To obtain the k-nearest neighbors of a test image descriptor, several kinds of distances can be considered. In this study, four distance measurements are implemented and compared. In the next lines, these distances are formalized. Considering $\vec{r} = \{r_i\}$, $i = 1, \dots, l$ and $\vec{s} = \{s_i\}$, $i = 1, \dots, l$, the two data vectors whose distance we want to obtain:

1. Weighted metric distance:

$$dist_p(\vec{r}, \vec{s}) = \left(\sum_{i=1}^l \omega_i \cdot |r_i - s_i|^p \right)^{\frac{1}{p}} \quad (6)$$

If we consider $\omega_i = 1, i = 1, \dots, l$, the Minkowski distance is obtained. Two particular cases will be considered: $dist_1$ (Manhattan distance), which is defined from the Minkowski distance with $p = 1$, and $dist_2$ (Euclidean distance), doing $p = 2$.

- Pearson correlation coefficient. It is a similitude coefficient that can be obtained as:

$$sim_{Pca}(\vec{r}, \vec{s}) = \frac{\vec{r}_d^T \cdot \vec{s}_d}{|\vec{r}_d| |\vec{s}_d|} \quad (7)$$

where $\vec{r}_d = [r_1 - \bar{r}, \dots, r_l - \bar{r}]$ and $\vec{s}_d = [s_1 - \bar{s}, \dots, s_l - \bar{s}]$, $\bar{r} = \frac{1}{l} \sum_j r_j$, $\bar{s} = \frac{1}{l} \sum_j s_j$. It takes values in the range $[-1, +1]$. From this similitude coefficient, a distance measure can be defined as:

$$dist_3(\vec{r}, \vec{s}) = 1 - sim_{Pca}(\vec{r}, \vec{s}) \quad (8)$$

- Inner product: It is also a similitude coefficient that can be calculated as the scalar product between the two vectors to compare.

$$sim_{cos}(\vec{r}, \vec{s}) = \frac{\vec{r}^T \cdot \vec{s}}{|\vec{r}| |\vec{s}|} \quad (9)$$

As shown in the equation, \vec{r} and \vec{s} are usually normalized. In this case, this measure is known as *cosine similitude* and takes values in the range $[-1, +1]$. The corresponding distance value is:

$$dist_4(\vec{r}, \vec{s}) = 1 - sim_{in}(\vec{r}, \vec{s}) \quad (10)$$

Therefore, the four distance measurements compared along this section are: $dist_1$ (Manhattan distance), $dist_2$ (Euclidean distance), $dist_3$ (Pearson correlation-based distance) and $dist_4$ (cosine similitude-based distance).

First, the success rate of each algorithm is studied. It assesses the ability of the localization algorithm to calculate correctly the first nearest neighbor (i.e., to identify correctly the position of the model which geometrically the nearest one to the test position). Figures 6–12 show the success rate, expressed on a per unit base. For comparative purposes, all the results are expressed in the same color scale.

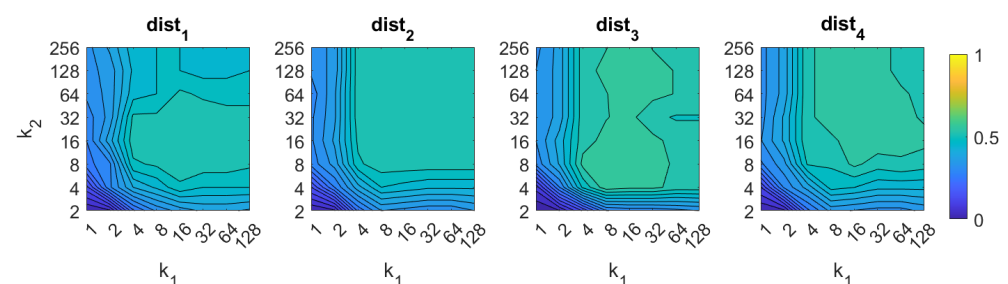


Figure 6. FS image retrieval problem. Success rate of the method. k_1 and k_2 are, respectively, the number of rows and columns of the descriptor (Table 1).

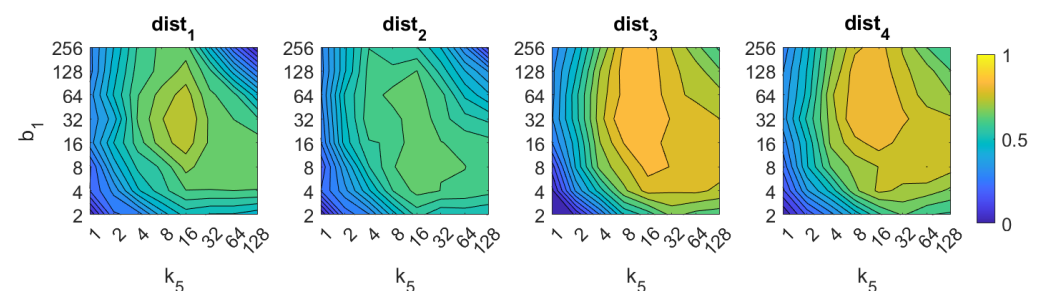


Figure 7. HOG image retrieval problem. Success rate of the method. k_5 is the number of horizontal cells and b_1 the number of bins per histogram (Table 1).

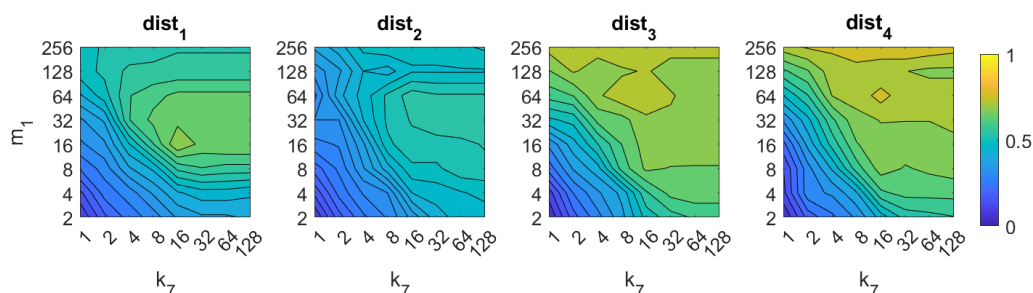


Figure 8. Gist image retrieval problem. Success rate of the method. k_7 is the number of horizontal blocks and m_1 the number of Gabor filters to build the descriptor (Table 1).

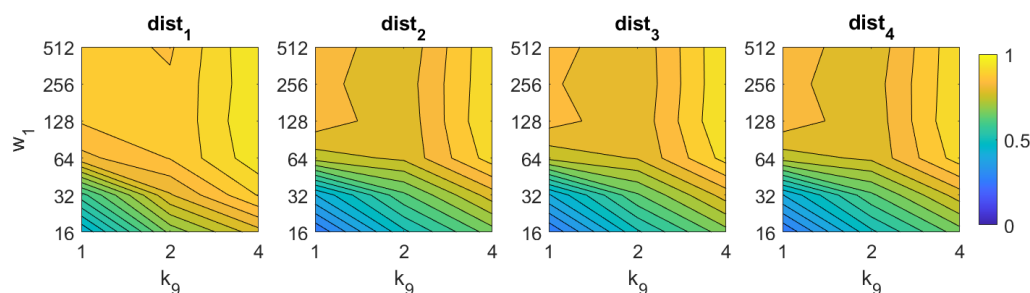


Figure 9. WS image retrieval problem. Success rate of the method. k_9 is the number of horizontal cells and w_1 the number of windows per cell (Table 1).

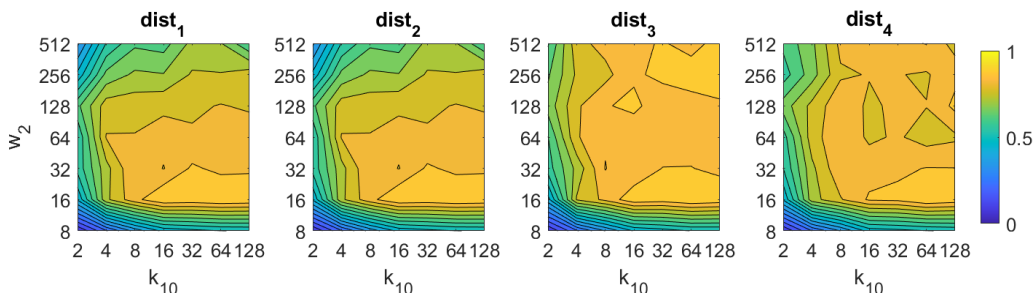


Figure 10. BG image retrieval problem. Success rate of the method. k_{10} is the number of horizontal cells and w_2 the number of windows per cell (Table 1).

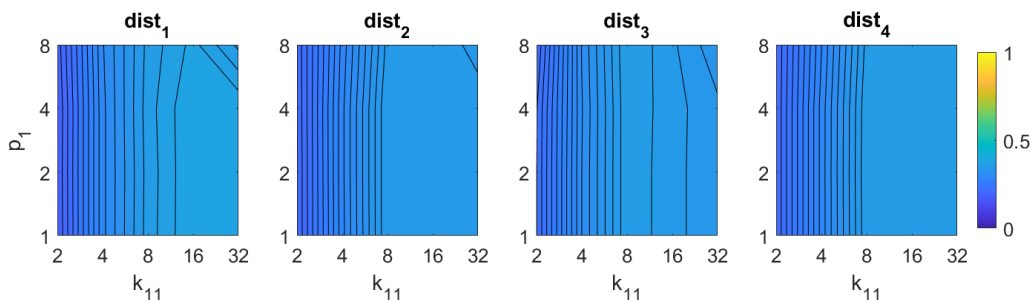


Figure 11. RT-F image retrieval problem. Success rate of the method. k_{11} is the number of blocks and p_1 the relative angle (deg) between the lines in each set (Table 1).

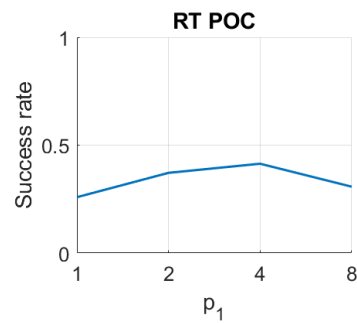


Figure 12. RT-POC image retrieval problem. Success rate of the method. p_1 is the relative angle (deg) between the lines in each set (Table 1).

The behavior of the FS changes slightly depending on the distance measurement used. The best results are obtained with $dist_3$ and $dist_4$ with an intermediate number of rows and an intermediate to high number of columns. In all cases, an excessively low number of rows and/or columns provides bad results. The best accuracy is 60%, and it is obtained with the distance $dist_3$ and $k_1 = k_2 = 8$.

About HOG, the best results are also obtained with distances $dist_3$ and $dist_4$. In both cases, the number of horizontal cells k_5 must be an intermediate value, around 16. A higher number does not improve the accuracy of the method. The number of bins per histogram b_1 must take values from intermediate to high, starting from 16. In the case of distances $dist_1$ and $dist_2$, an excessively high number of cells and bins also provides remarkably bad results. The best accuracy is 89%, and it is obtained with the distance $dist_3$ and $k_5 = 8, b_1 = 32$.

In the case of *gist*, the best results are obtained again using the distances $dist_3$ and $dist_4$. In these cases, the accuracy increases as the number of masks m_1 does. It is not necessary a high number of masks m_1 to obtain good results. The best accuracy is 89%, and it is obtained with the distance $dist_3$ and $k_7 = 32, m_1 = 256$.

In the case of *Wi-SURF*, the best results are obtained using the distances $dist_1$ and $dist_3$. In these cases, the image retrieval problem is solved with a better rate when using high values of k_9 (around 4). The process performs correctly with intermediate and high number of windows per cell w_1 , starting from 128. The best rate is 97%, and it is obtained with the distance $dist_1$ and $k_9 = 4, w_1 = 512$.

If we analyze now *BRIEF-gist*, the best results are obtained using the distances $dist_3$ and $dist_4$. A high number of horizontal cells k_{10} is needed to obtain suitable results, about 64. A high number of windows w_2 does not improve the results necessarily, but remarkably bad results are obtained using low values of k_{10} or w_2 . The best accuracy obtained with *BG* is 93%, and it is obtained with the distance $dist_3$ and $k_{10} = 64, w_2 = 16$.

Finally, in the case of *RT*, the results are not competitive if they are compared with the rest of the descriptors. On the one hand, using the Radon transform along with the Fourier Signature, the best results are obtained with the distances $dist_1$ and $dist_4$. In this case the parameters have less relevance on the results, but in general, high values of k_{11} and low values of p_1 lead to better rates. Using *RT-F*, the best accuracy is 39%, and it is obtained with the distance $dist_1$ and $k_{11} = 32, p_1 = 1$. On the other hand, using the POC method, the best rate is 41% obtained with $p_1 = 4$.

Analyzing globally these figures, *Wi-SURF* is the description algorithm that presents the best absolute success rate, when it is used along with $dist_1$. In general, the distance $dist_3$ performs much better than the rest in almost all the cases. *HOG*, *gist* and *BRIEF-gist* are also acceptable methods. Taking into account the challenging characteristics of the environment, they provide remarkably good results.

Apart from the success rate, it is also worth studying the computational cost of the process, to evaluate whether the localization task could be carried out in real time. Figures 13–19 show the necessary time to obtain the nearest neighbor, depending on the size of the position descriptor. The average value after all the experiments is shown,

expressed in seconds. A logarithmic scale has been used to represent efficiently the time in the color scale.

The experiments have been carried out with a CPU Intel Core i7-9700 at 3 GHz and using the mathematical tool Matlab. These time results are not absolute, they depend of the computer which runs the process. They are comparable because all the calculations have been done with the same machine.

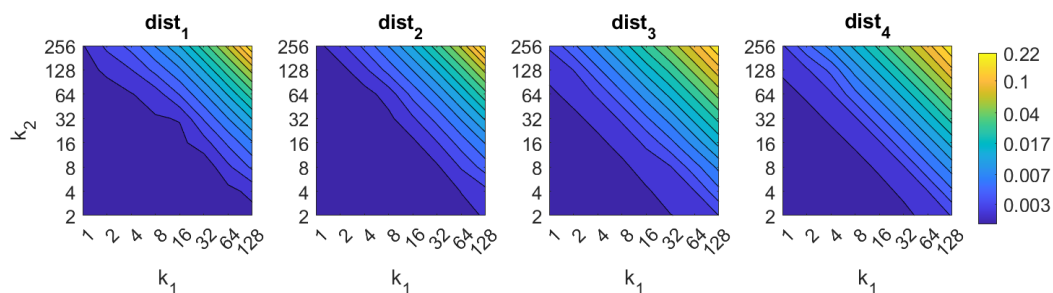


Figure 13. FS image retrieval problem. Computational time.

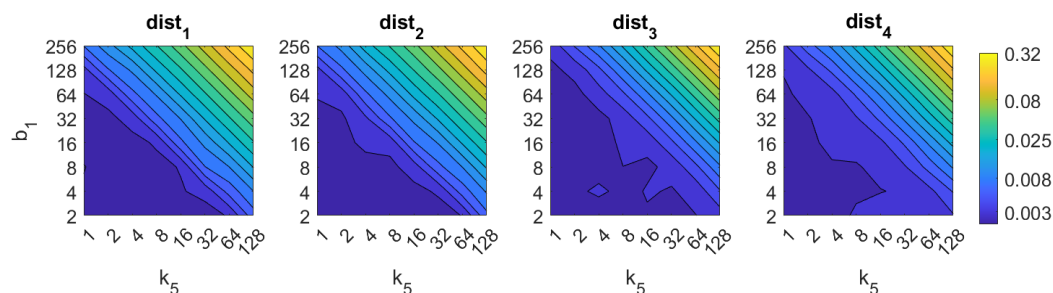


Figure 14. HOG image retrieval problem. Computational time.

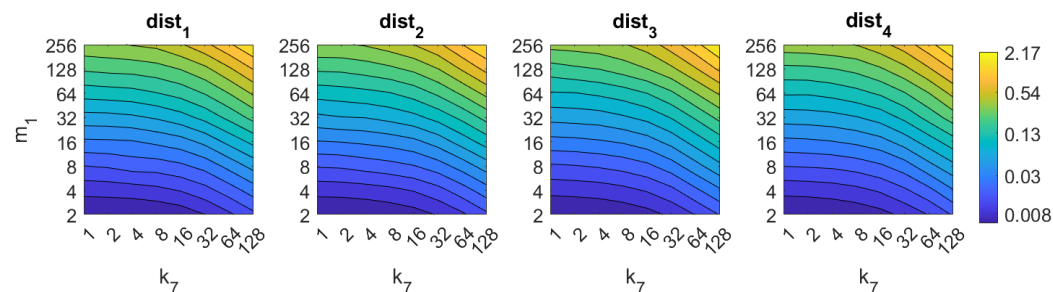


Figure 15. Gist image retrieval problem. Computational time.

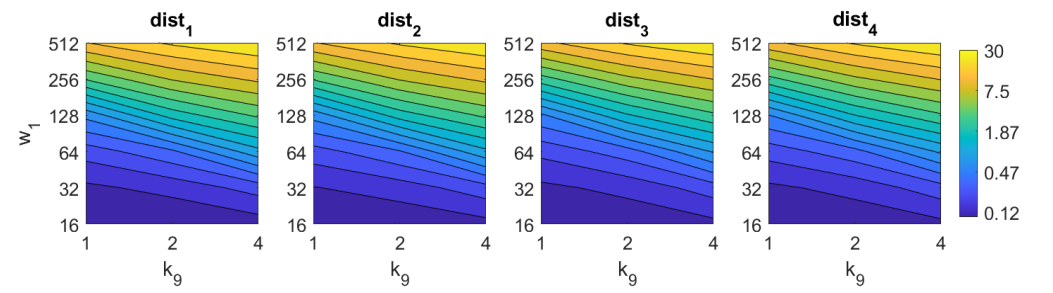


Figure 16. WS image retrieval problem. Computational time.

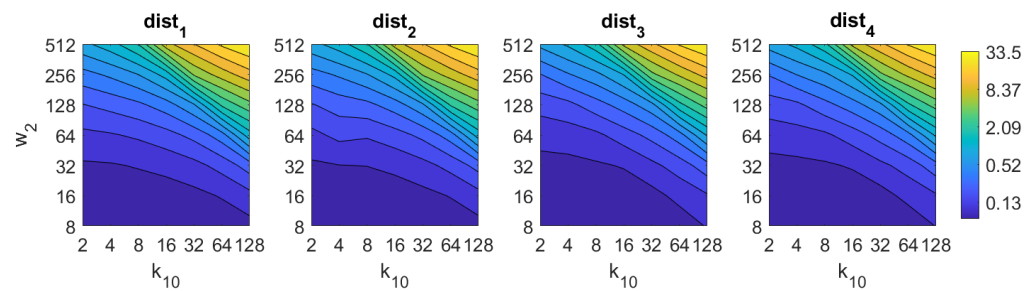


Figure 17. BF image retrieval problem. Computational time.

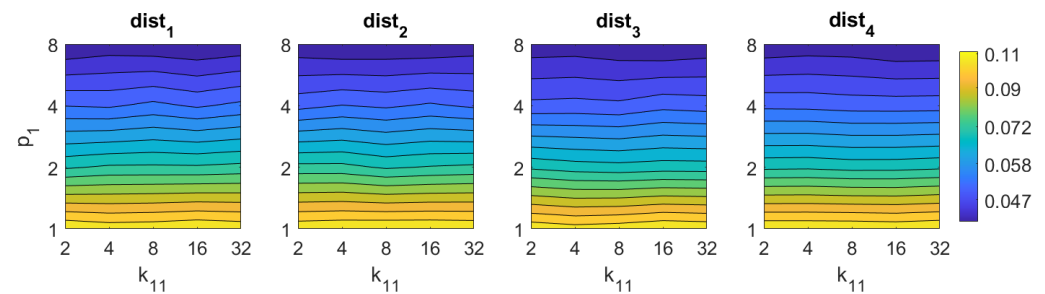


Figure 18. RT-F image retrieval problem. Computational time.

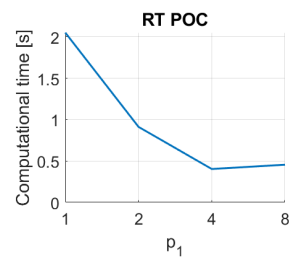


Figure 19. RT-POC image retrieval problem. Computational time.

First, FS is the quicker algorithm. The average time per *test image* is under 0.02 s for the majority of configurations. Only when both k_1 and k_2 take high values, the computational time takes values around 0.22 s. Both parameters have a similar influence on the computational cost. Second, the computational cost of HOG is slightly higher than FS, depending on the configuration of the parameters. Both parameters b_1 and k_3 have similar influence on this time. When their values are high it is possible to find some results where the runtime takes around 0.32 s. Third, *gist* is computationally more an expensive algorithm. m_1 has a strong influence on the necessary time. A high number of masks along with high values of k_7 make the time per image to take values around 2.1 s. Anyway, it is possible to find configurations that provide acceptable computational times with a lower number of components.

The second group of descriptors, in which each descriptor should be shifted until finding the relative orientation before retrieving the image, are considerably slower. On the one hand, *Wi-SURF* needs more than 2 s with most of the configurations. w_1 has more influence on the computational time so, as far as possible, it is better to avoid high values of this parameter. High values of the parameters can lead to times up to 30 s. On the other hand, *Wi-SURF* is the computationally most expensive method. w_2 has a strong influence on the process, and produces times about 33.5 s.

Finally, the method based on the Radon transform and Fourier performs quickly, with times typically under 0.1 s. The method based on Radon transform and POC leads to times around 0.5 s with some configurations of p_1 . Notwithstanding that, since the descriptors based on the Radon transform have proved to perform poorly in the image retrieval task, these descriptors are not included in subsequent analyses.

5.2. Estimation of the Position

The second set of experiments assesses the ability of each description method to estimate correctly the position of the robot, when noise or occlusions are present, depending on the size of the descriptor and the type of distance considered.

For each test image, the position descriptor is obtained and compared with all the position descriptors in the map. The 1st nearest neighbor is then retained, using any distance measurement. In the cases that it is possible, a k-d tree has been implemented to make efficiently this search. After obtaining the nearest neighbor, the Euclidean distance between the real position of the robot at time instant t and the position of the nearest neighbor is considered the position error.

Figures 20 and 21 present the results obtained with the Fourier Signature considering the presence of noise or occlusions, respectively, in the test images. In these figures, first, no filter is considered (a) and second, a homomorphic filtering is carried out both to the reference and the test images (b). The result is expressed then as the average position error, expressed in cm after considering the 1232 test images. The horizontal axis expresses the percentage of information considered per configuration, expressed in logarithmic scale. The ticks of each graphical representation are $\{2^{-15}, 2^{-14}, 2^{-13}, \dots, 2^{-2}, 2^{-1}\}$ which correspond, respectively, to the next percentages of information $\{0.003\%, 0.06\%, 0.12\%, \dots, 25\%, 50\%\}$. These percentages express the information contained in each descriptor with respect to the information contained in each original panoramic image $\left(\frac{k_1 \cdot k_2}{N_1 \cdot N_2} \cdot 100\right)$. In general, the use of homomorphic filtering worsens the results. As expected, the higher the level of noise, the higher the error. However, $dist_1$ and $dist_2$ present a more robust behavior when noise is present. About the presence of occlusions, the FS descriptor is quite sensitive to this phenomenon and the results worsen substantially when the percentage of occlusion increases.

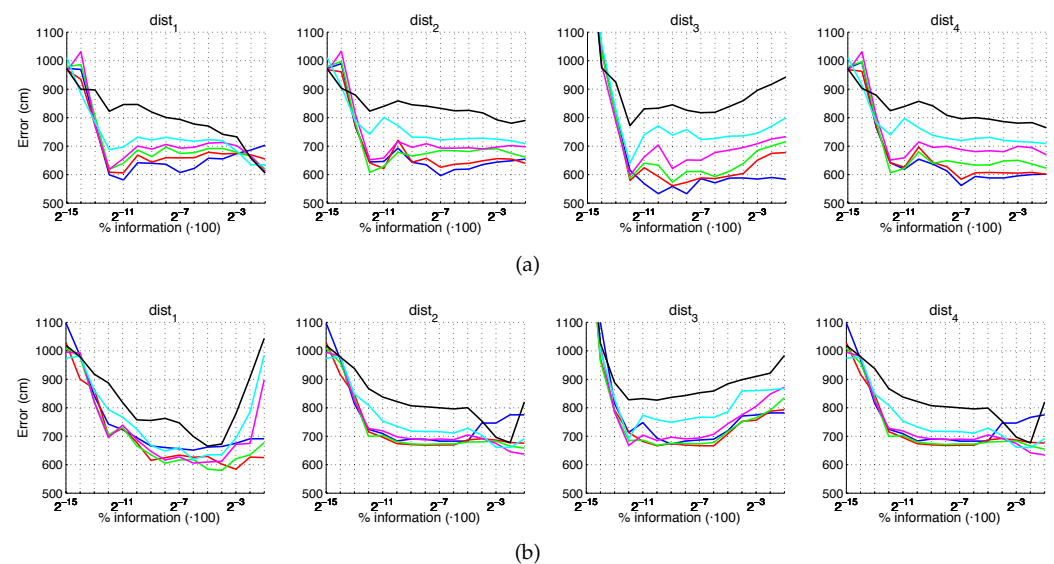


Figure 20. FS average localization error with noise: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Noise 1, — Noise 2, — Noise 3, — Noise 4, — Noise 5.

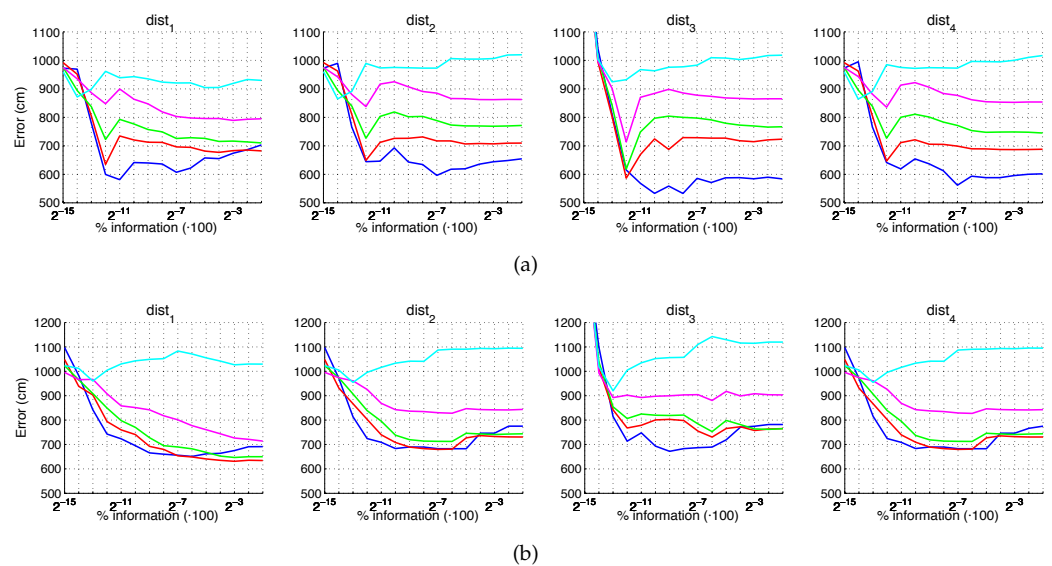


Figure 21. FS average localization error with occlusions: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Occlusion 1, — Occlusion 2, — Occlusion 3, — Occlusion 4.

Figures 22 and 23 present the results obtained with the Histogram of Oriented Gradients considering the presence of noise or occlusions, respectively, in the test images. In these figures, first, no filter is considered (a) and second, a homomorphic filtering is carried out both to the reference and the test images (b). Like in the case of FS, the percentages in the horizontal axis express the information contained in each descriptor with respect to the information contained in each panoramic image. In the case of HOG they can be obtained as $\left(\frac{k_5 \cdot b_1}{N_1 \cdot N_2} \cdot 100\right)$. In presence of noise, the use of homomorphic filtering only improves the results with distances $dist_3$ and $dist_4$ and with low level of noise. Intermediate percentages of information tend to present the best absolute results so it is not necessary to store a big quantity of information during the construction of the descriptor. In presence of noise, the best absolute results are obtained with $dist_3$, no filter and intermediate quantity of information. Comparing to the other description methods, HOG stands out thank to its robustness against presence of occlusions in the test images.

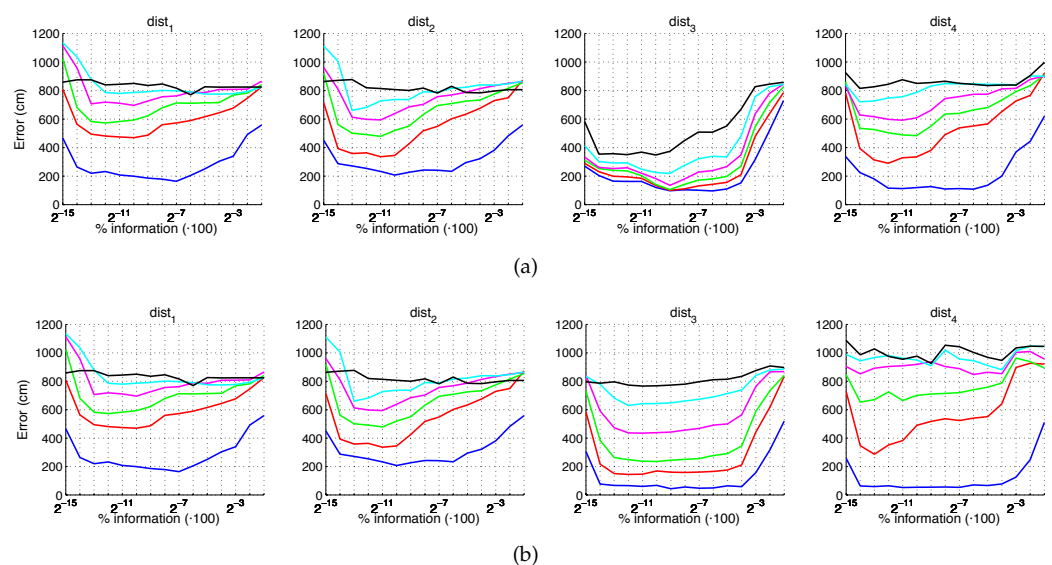


Figure 22. HOG average localization error with noise: (a) no filter and (b) homomorphic filter. Legend: — Original, — Noise 1, — Noise 2, — Noise 3, — Noise 4, — Noise 5.

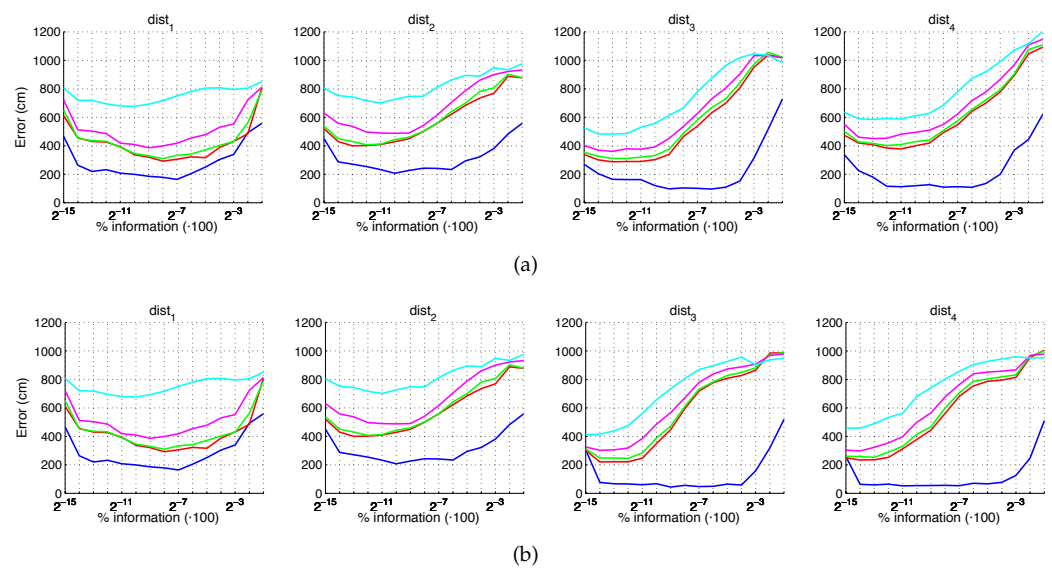


Figure 23. HOG average localization error with occlusions: (a) no filter and (b) homomorphic filter. Legend: — Original, — Occlusion 1, — Occlusion 2, — Occlusion 3, — Occlusion 4.

Additionally, Figures 24 and 25 present the results obtained with *gist* considering the presence of noise or occlusions, respectively, in the test images. In these figures, first, no filter is considered (a) and second, a homomorphic filtering is carried out both to the reference and the test images (b). Like in the case of FS, the percentages of information contained in each descriptor with respect to the information contained in each panoramic image can be obtained as $\left(\frac{2 \cdot k_7 \cdot m_1}{N_1 \cdot N_2} \cdot 100\right)$. The use of homomorphic filtering does not improve the localization results in any case. In the presence of noise, *dist*₃ presents the best results when considering an intermediate percentage of information.

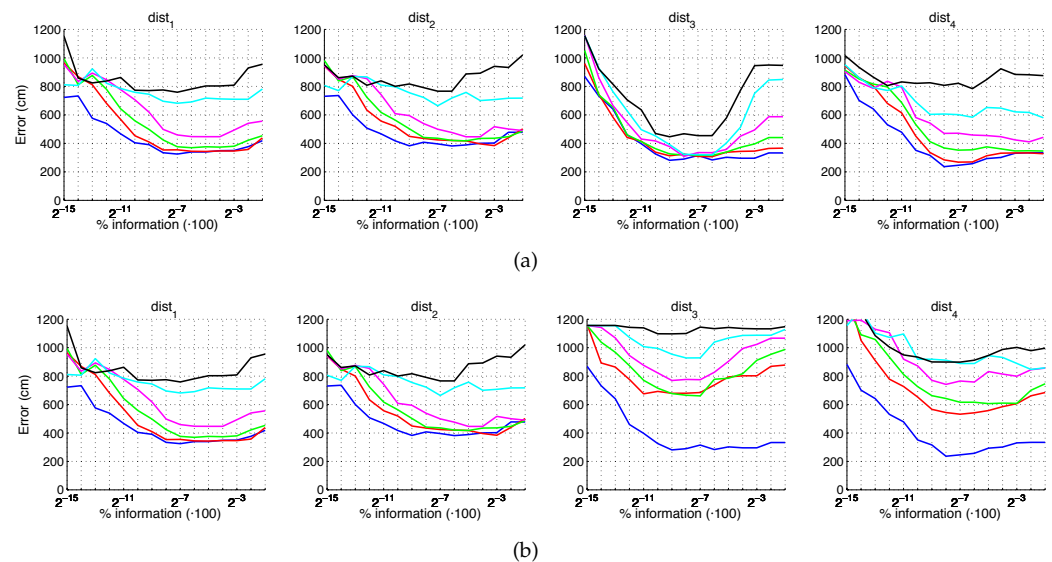


Figure 24. *Gist* average localization error with noise: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Noise 1, — Noise 2, — Noise 3, — Noise 4, — Noise 5.

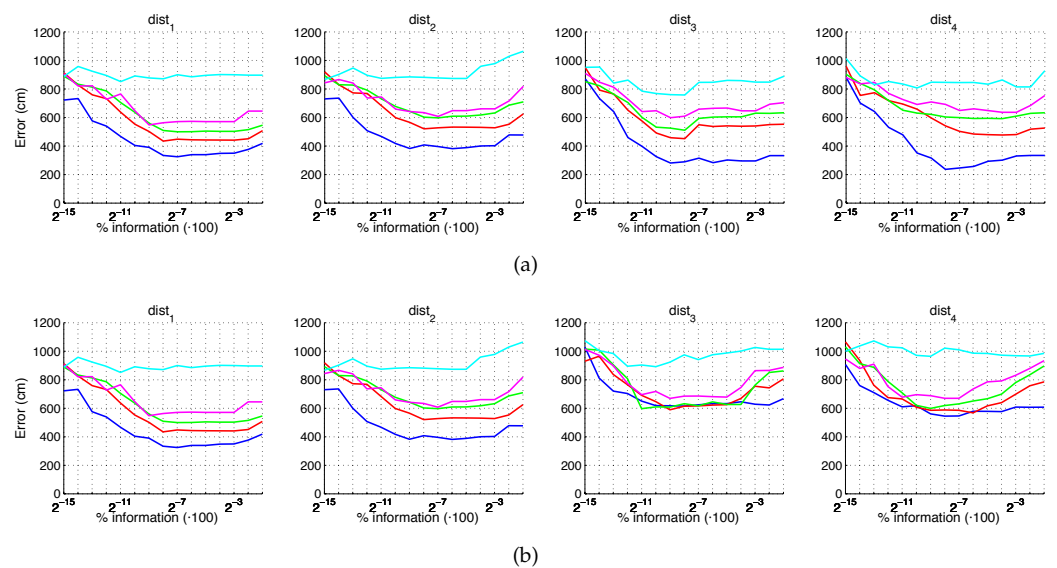


Figure 25. Gist average localization error with occlusions: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Occlusion 1, — Occlusion 2, — Occlusion 3, — Occlusion 4.

Fourthly, Figures 26 and 27 present the results obtained with *Wi-SURF* considering the presence of noise or occlusions, respectively, in the test images. In these figures, first, no filter is considered (a) and second, a homomorphic filtering is carried out both to the reference and the test images (b). The information contained in each descriptor with respect to the information contained in each panoramic image can be obtained as $\left(\frac{k_g \cdot w_1 \cdot 64}{N_1 \cdot N_2} \cdot 100\right)$. The use of homomorphic filtering does not reduce the localization error. In this case, the performance of the descriptor is severely influenced by the presence of noise. It is very significant that results without noise and occlusion are better than the errors obtained with the previous descriptors, but when these effects appear on the scene the results worsen sharply. In general, $dist_1$ and $dist_3$ present the best results when considering an intermediate or high percentage of information.

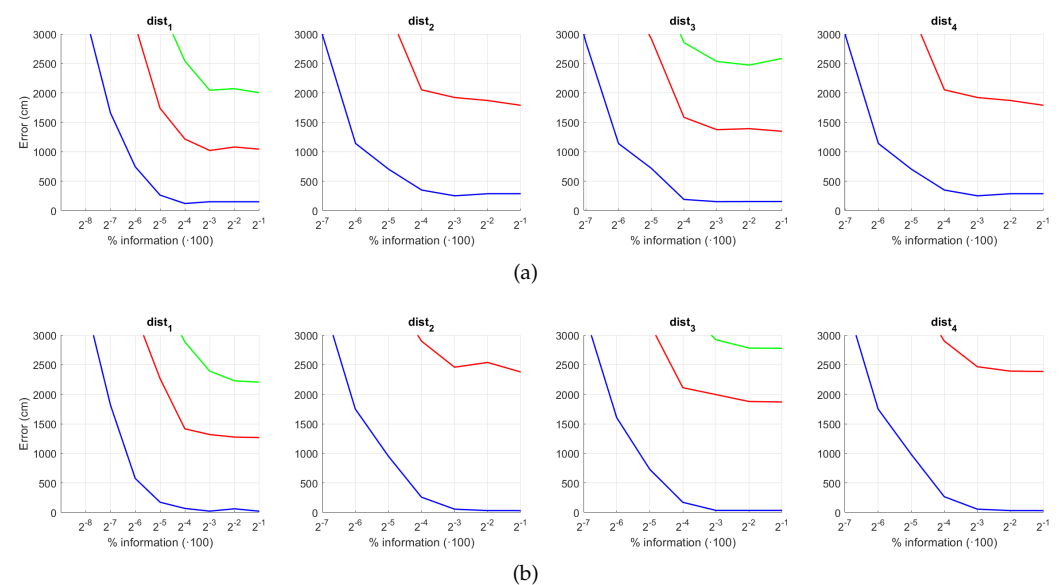


Figure 26. WS average localization error with noise: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Noise 1, — Noise 2, — Noise 3, — Noise 4, — Noise 5.

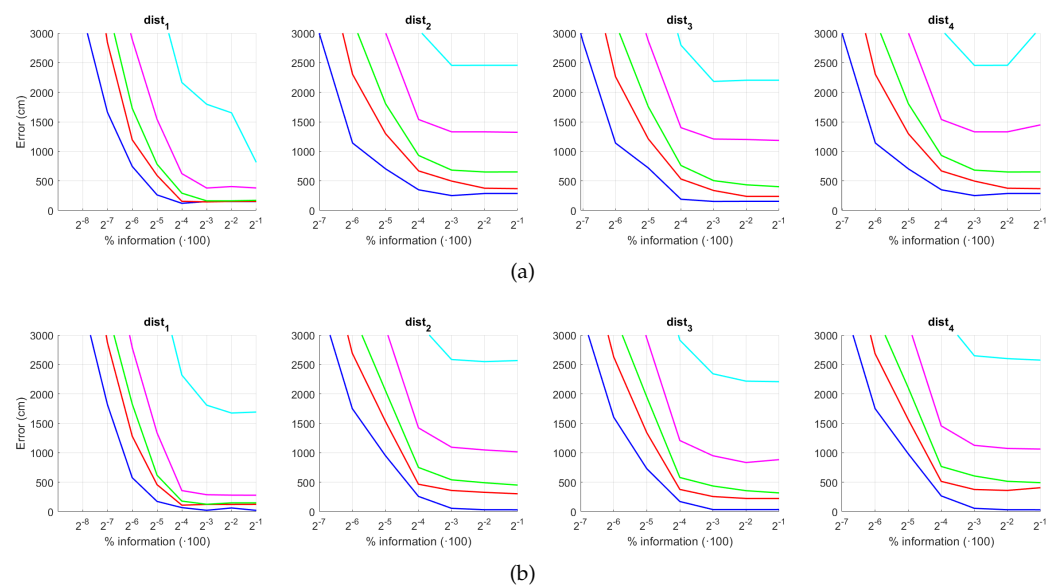


Figure 27. WS average localization error with occlusions: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Occlusion 1, — Occlusion 2, — Occlusion 3, — Occlusion 4.

Figures 28 and 29 present the results obtained with the *BRIEF-gist* method considering the presence of noise or occlusions, respectively, in the test images. In these figures, first, no filter is considered (a) and second, a homomorphic filtering is carried out both to the reference and the test images (b). Like in previous figures, the percentages in the horizontal axis express the information contained in each descriptor with respect to the information contained in each panoramic image. In the case of BG, they can be obtained as $\left(\frac{k_{10} \cdot w_2}{N_1 \cdot N_2} \cdot 100\right)$. In this case, the best results are achieved with an intermediate amount of information, so it is not necessary to store a big quantity of information when building the descriptors. In addition, in general terms, the filter tends to improve the results. Comparing to the other description methods, *BRIEF-gist* presents higher error in ideal conditions, but it controls its error when noise appears on the scenes, obtaining good results even with high quantity of noise. Additionally it performs correctly when no occlusions take part on the image but it works wrongly when this phenomenon appears.

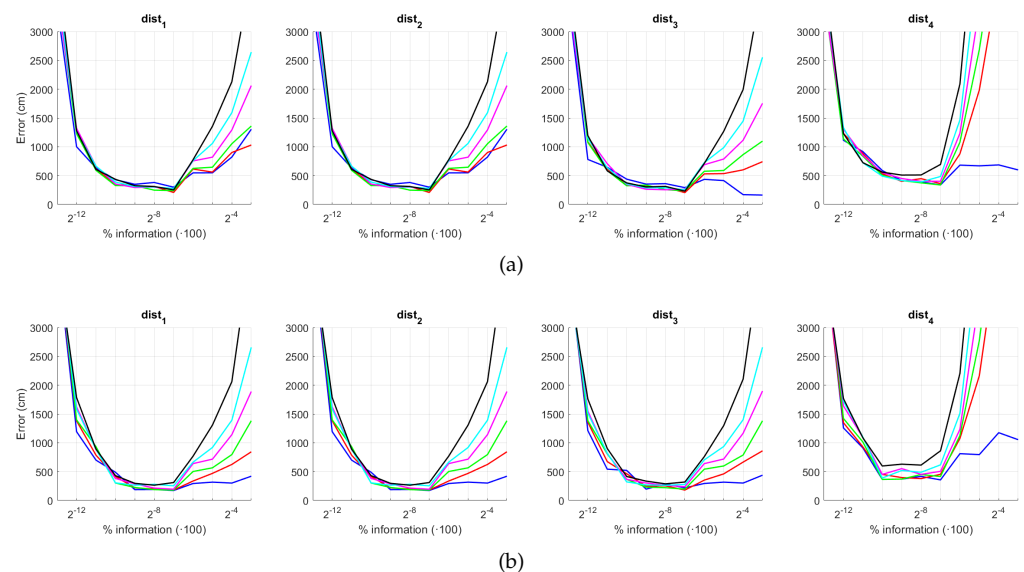


Figure 28. BG average localization error with noise: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Noise 1, — Noise 2, — Noise 3, — Noise 4, — Noise 5.

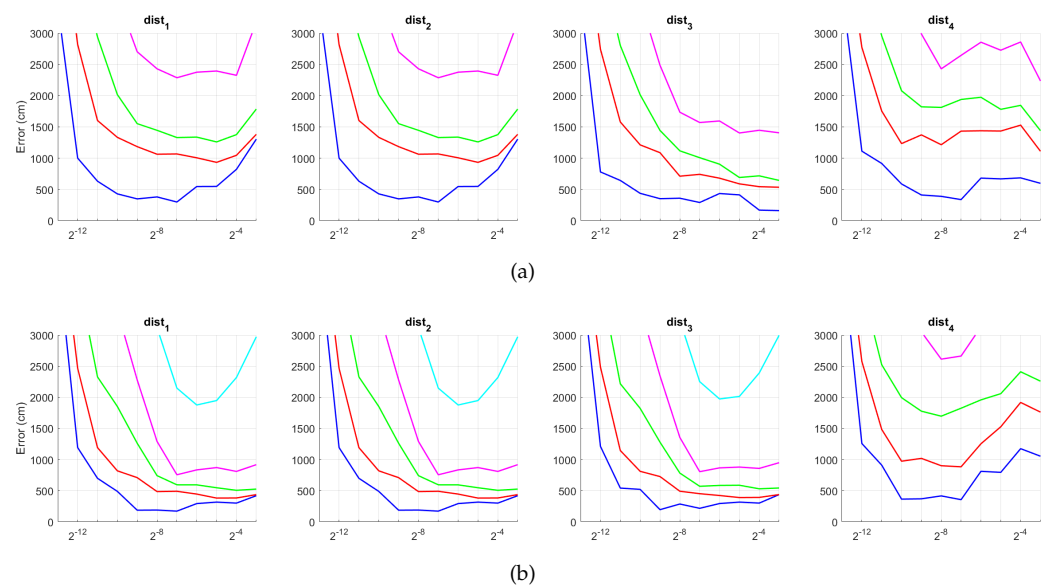


Figure 29. BG average localization error with occlusions: (a) no filter and (b) homomorphic filtering. Legend: — Original, — Occlusion 1, — Occlusion 2, — Occlusion 3, — Occlusion 4.

If we analyze jointly these results, we can arrive to some general conclusions. First, HOG presents very good localization results under ideal conditions. These results degrade in the presence of noise or occlusions, but some configurations resist these effects. Second, gist with no filter leads to worse results in ideal conditions, but it is robust against adverse effects, mainly against noise. Third, WS along with filter provides the best absolute localization results in ideal conditions. However, its performance sharply worsens with noise and occlusions. Finally, the results of BG in ideal conditions are not remarkable. However, this is the descriptor that presents more robustness in the presence of noise and occlusions, even in very unfavorable conditions.

5.3. Estimation of the Orientation

In this section, the problem of orientation estimation is addressed. To assess the performance of each description method in this task, independently of the results of the position estimation, the test image orientation descriptor is always compared with the orientation descriptor of the map image which was captured in the geometrically closest position. The problem is solved using the algorithms presented in Section 3, except those based on Radon transform, which proved to perform poorly in the image retrieval task.

First, the results obtained with the Fourier Signature are presented. Figure 30 shows the results of the orientation estimation. The influence of noise is also assessed in this figure. The results are expressed as average orientation error, in degrees, after repeating the experiment with the 1232 test images. This figure shows that the algorithm is very robust against the presence of noise. The optimal configuration is an intermediate to high number of rows (k_3) and an intermediate number of columns (k_4). A high number of columns worsens the results. Additionally, the presence of occlusions in the orientation estimation process is assessed in Figure 31. This figure shows that the influence of occlusions is higher, since the results tend to worsen as the level of occlusion increases. Nevertheless, some configurations of the parameters permit obtaining an average error lower than 10 deg even with 40% occlusions. The computational time of the orientation estimation process is shown in Figure 32, expressed in seconds. The descriptor based on FS is able to estimate the orientation relatively quickly for most configurations of k_3 and k_4 and only high values of both parameters produce a relatively high computation time.

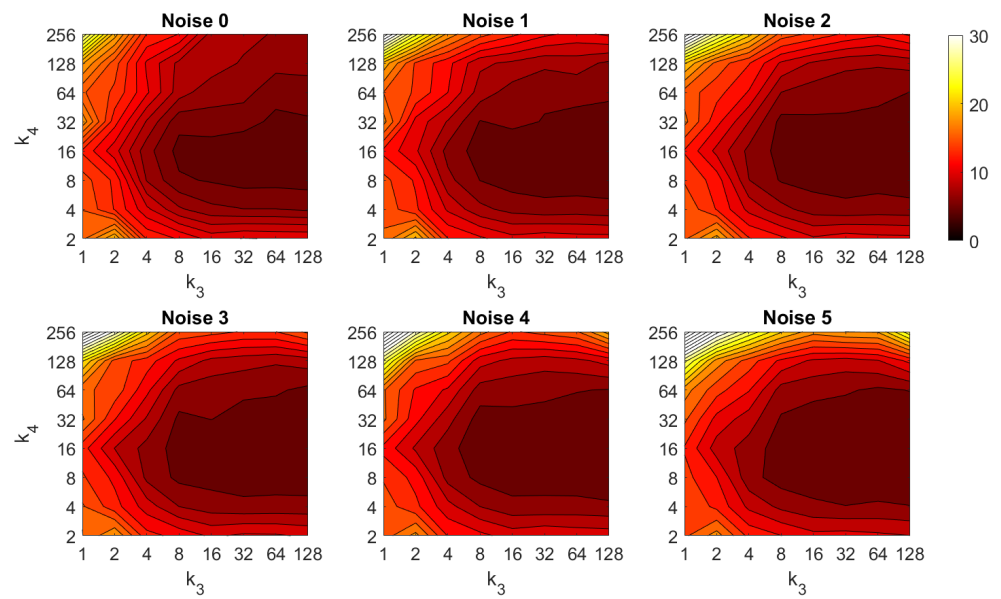


Figure 30. FS orientation estimation in the presence of noise. Average orientation error (deg).

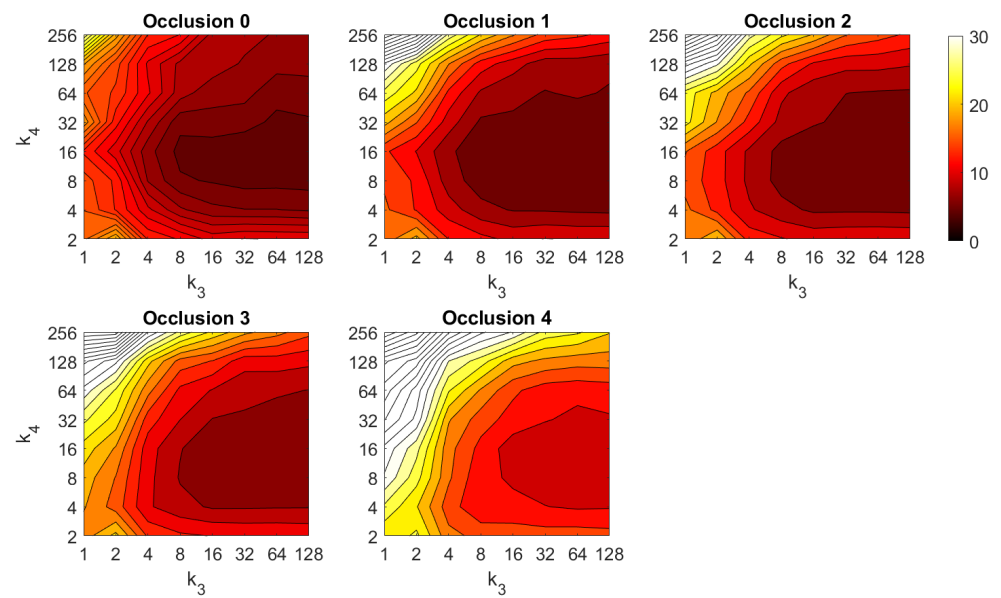


Figure 31. FS orientation estimation in the presence of occlusions. Average orientation error (deg).

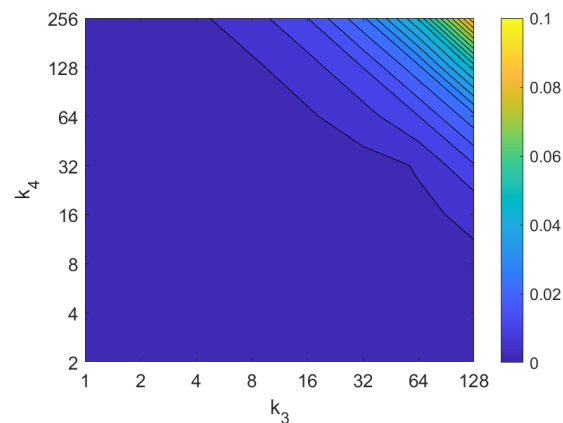


Figure 32. FS orientation estimation. Average computation time (s).

Second, the performance of the HOG descriptor is assessed, considering several values of the parameters l_1 (width of the vertical cells in the orientation descriptor) and d_1 (distance between consecutive vertical cells, which are overlapped). Figure 33 shows the average orientation error after considering all the test images. In addition, the influence of the presence of different levels of noise in the test images is analyzed. In general terms, low to intermediate values of d_1 and high values of l_1 produce the best results (lower orientation error). In addition, HOG proves to be a descriptor which is robust against the presence of noise, since the results do not change substantially as the level of noise increases. In general, HOG tends to present better results in orientation estimation comparing with FS. Furthermore, the influence of partial occlusions in orientation estimation is shown in Figure 34. As with FS, the influence of occlusions in the orientation estimation is substantial, and the results degrade quickly as the percentage of occlusions increases. Notwithstanding that, high values of l_1 tend to produce relatively low orientation error, independently of the level of occlusions. Finally, Figure 35 shows the necessary time to estimate the orientation (average time, expressed in seconds, after considering all the test images). Most configurations of l_1 and d_1 produce a relatively low computation time. Only very high values of l_1 combined with low values of d_1 output a substantially high calculation time.

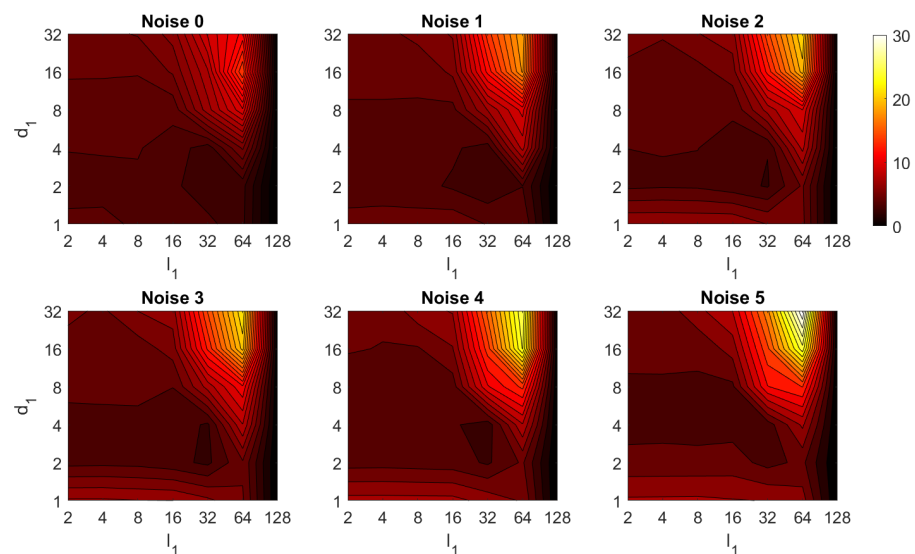


Figure 33. HOG orientation estimation in the presence of noise. Average orientation error (deg).

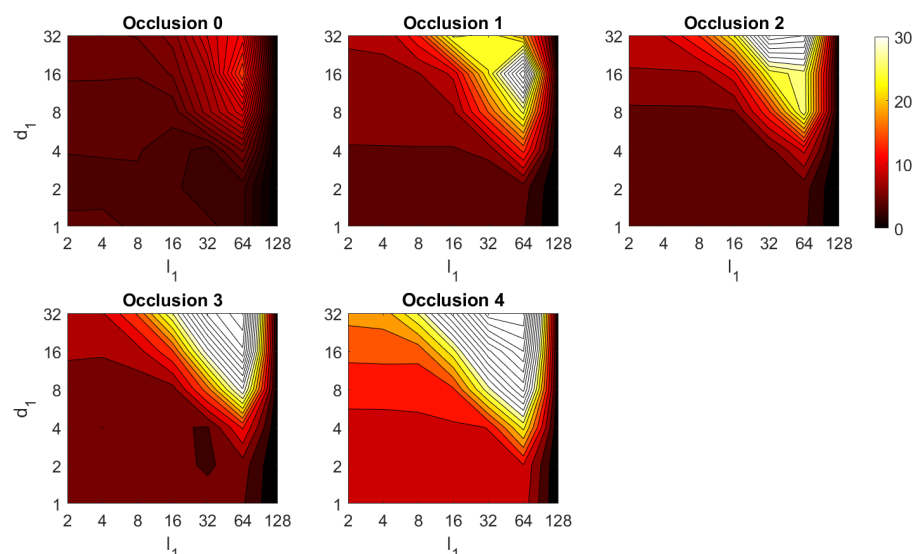


Figure 34. HOG orientation estimation in the presence of occlusion. Average orientation error (deg).

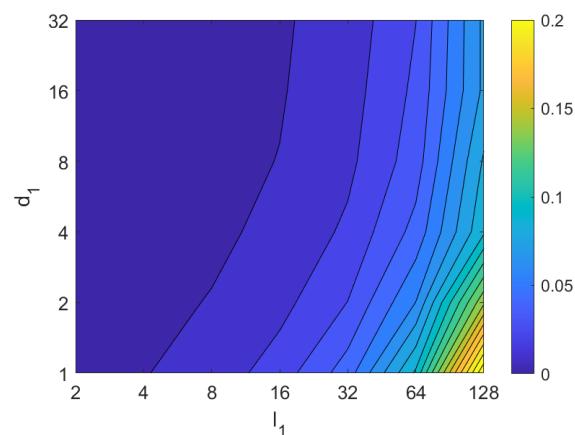


Figure 35. HOG orientation estimation. Average computation time (s).

Third, the results of relative orientation estimation with *gist* are presented and commented. Figure 36 shows the average orientation error (degrees) when considering different configurations of l_2 (width of the vertical blocks in the orientation descriptor) and d_2 (distance between two consecutive vertical blocks, which are overlapped). The influence of the level of occlusions is checked in this figure. In the case of this description method, the orientation error tends to increase as d_2 does. However, as in the case of HOG, high values of the width of the vertical blocks produce relatively good results independently of the value of d_2 . To finish the experiments, the necessary time to estimate the orientation (average time in seconds) is shown in Figure 37. The figure shows that d_2 is the parameter that has a predominant influence upon the calculation time. Low values of this parameter produce a comparatively high computation time.

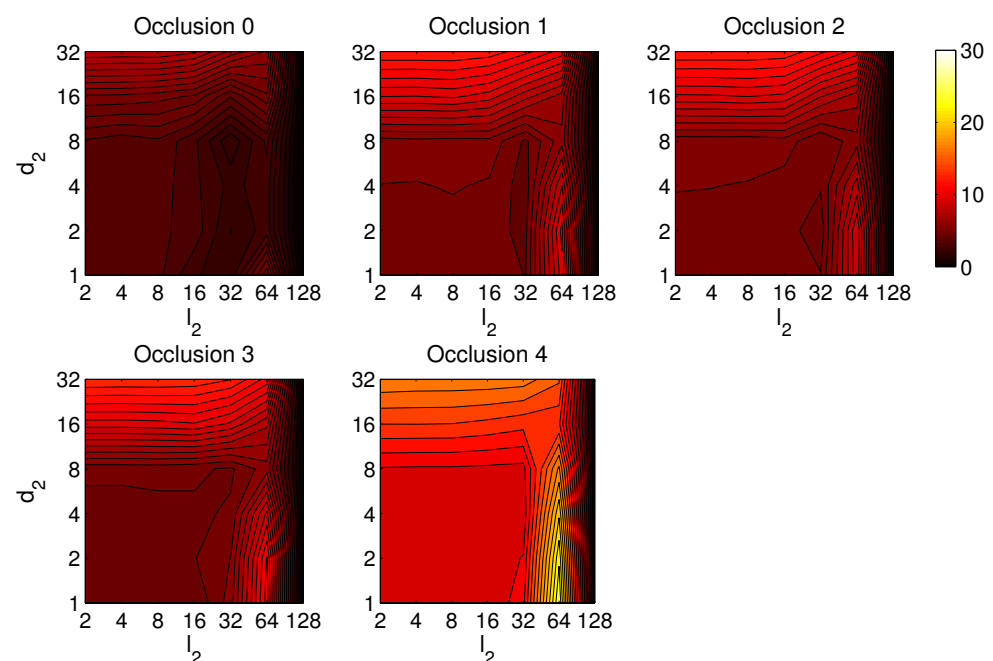


Figure 36. Gist orientation estimation in the presence of occlusion. Average orientation error (deg).

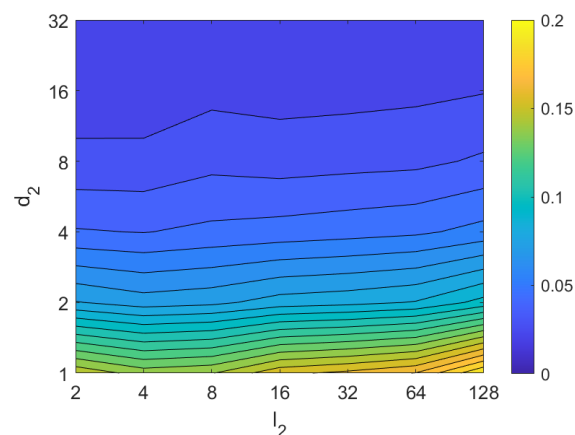


Figure 37. Gist orientation estimation. Average computation time (s).

In addition, the results of relative orientation estimation with *Wi-SURF* are presented and commented. Figure 38 shows the average orientation error (degrees) taking into account the noise influence considering the variation on the parameters k_g and w_1 . It shows a strong influence of the noise in the result. It is possible to check that results without noise are acceptable (about 5–10 deg), but the error increase considerably when the noise appears on the scenes. If the image is corrupted with noise with variance higher than $\sigma^2 = 0.0025$, the error is always more than 30 deg. The influence of the level of occlusions can be checked in the Figure 39. In the case of the occlusions, the results show more robustness, except for the results with 40% of occlusion that are considerably bad comparing with HOG. In general, the error tends to be optimized with middle values of w_1 . To finish the experiments, the necessary time to estimate the orientation (average time in seconds) is shown in Figure 40. The figure shows that w_1 is the parameter that has a predominant influence upon the calculation time. High values of this parameter produce a comparatively high computation time.

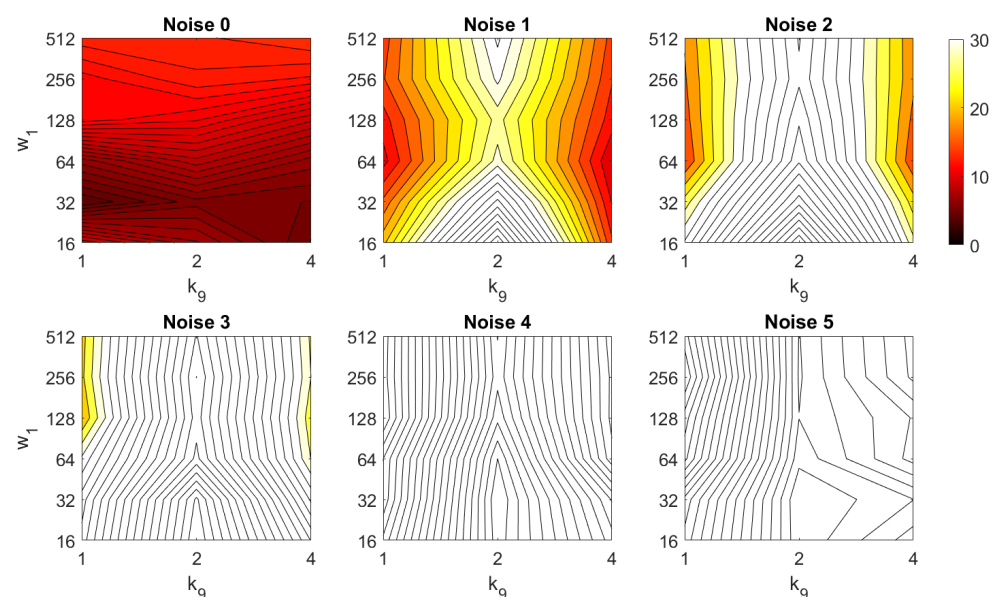


Figure 38. WS orientation estimation in the presence of noise. Average orientation error (deg).

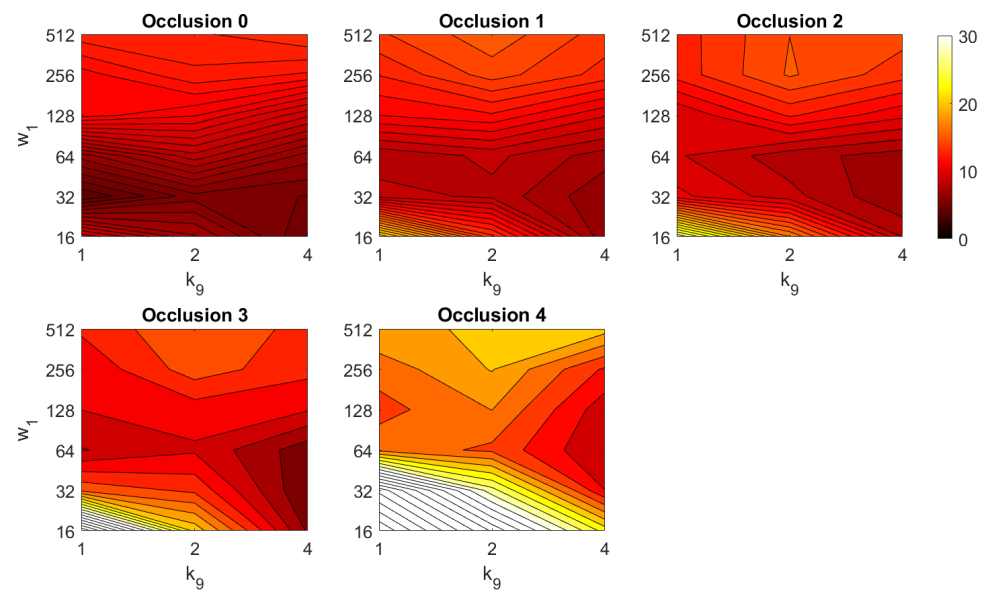


Figure 39. WS orientation estimation in the presence of occlusion. Average orientation error (deg).

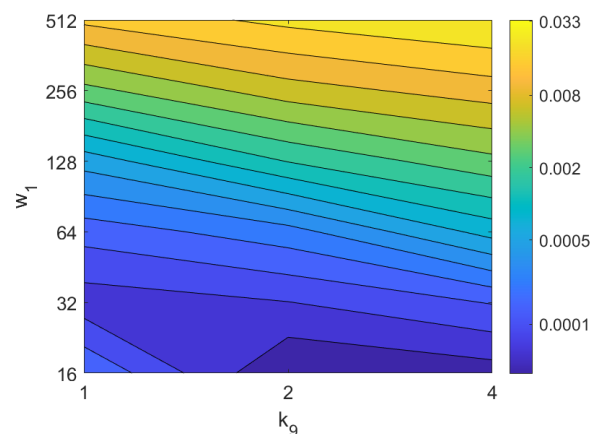


Figure 40. WS orientation estimation. Average computation time (s).

Additionally, the performance of the *BRIEF-gist* descriptor is assessed, considering several values of the parameters w_2 and k_{10} . Figure 41 shows the average orientation error after considering all the test images and the influence of the presence of different levels of noise. In general terms, the optimal configuration is an intermediate to high number of cells (k_{10}) and an intermediate number of windows (w_2). A high number of windows lead to worse results. In addition, *BRIEF-gist* proves to be a descriptor which is robust against the presence of noise, since the results do not change substantially as the level of noise increases. In general, *BRIEF-gist* tends to present better results in orientation estimation comparing with other descriptors. However, the influence of partial occlusions in orientation estimation has a worse influence, as shown in Figure 42. As with *Wi-SURF*, the algorithm performs considerably bad under the influence of occlusions. As before, intermediate values of w_2 output the best results. Finally, Figure 43 shows the necessary time to estimate the orientation. Most configurations of w_2 and k_{10} produce a relatively low computation time. Only very high values output a substantially high calculation time.

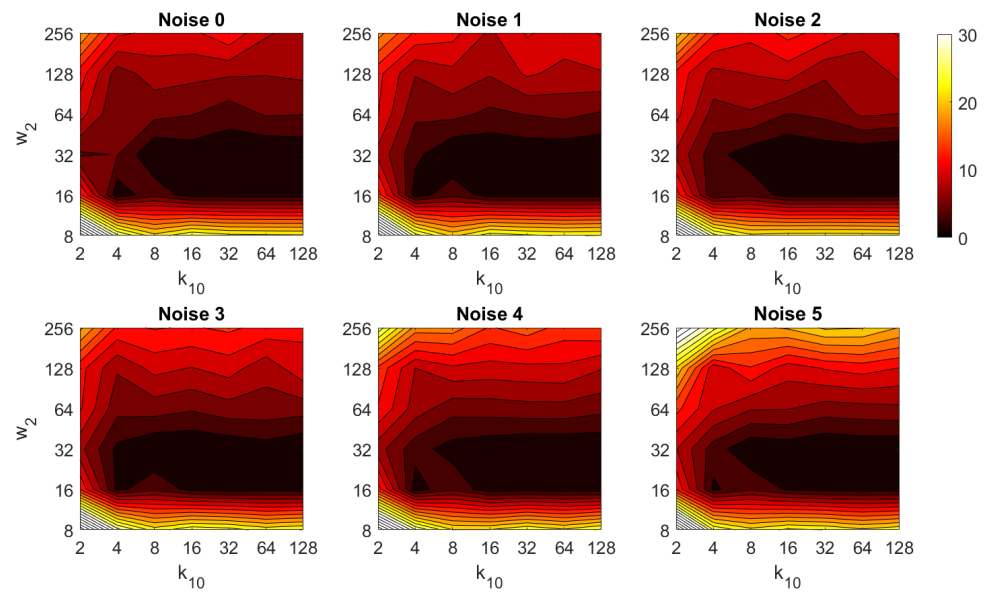


Figure 41. BG orientation estimation in the presence of noise. Average orientation error (deg).

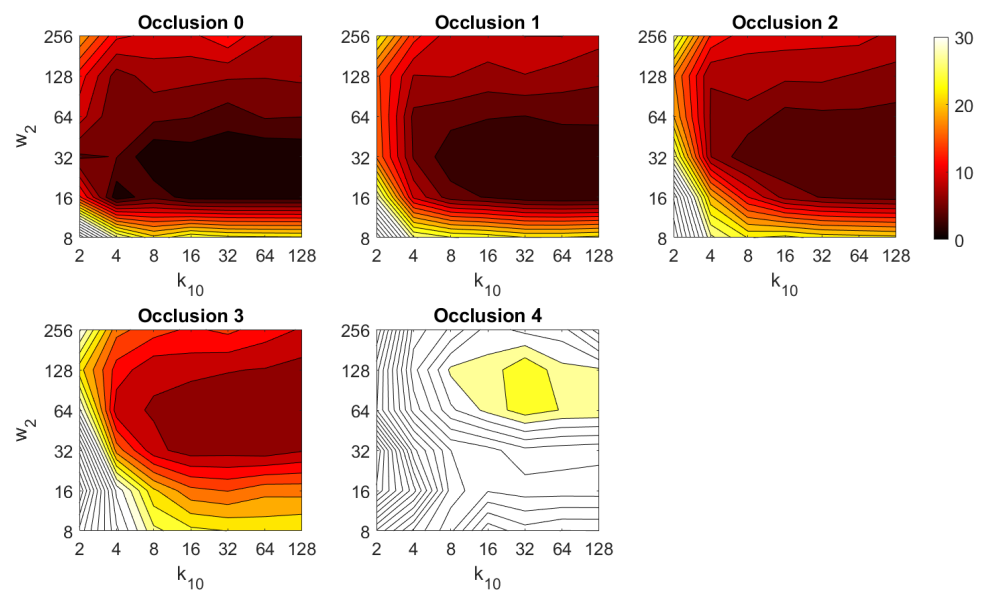


Figure 42. BG orientation estimation in the presence of occlusion. Average orientation error (deg).

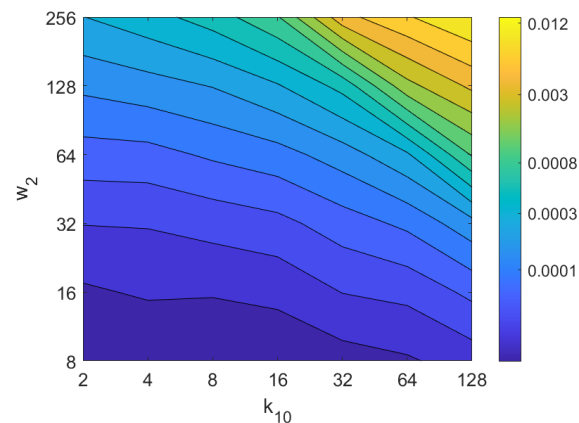


Figure 43. BG orientation estimation. Average computation time (s).

In general terms, HOG and *gist* produce relatively better results in the estimation of relative orientation, and the previous figures prove that it is possible to find some configurations of the most relevant parameters that offer a good balance between error and calculation time. Moreover, *Wi-SURF* and *BRIEF-gist* also offer acceptable errors in ideal conditions, and the calculation times are low. However, with these two descriptors, the orientation error tends to increase remarkably with the presence of occlusions and noise.

5.4. Evaluation with a Trajectory Dataset

To conclude the experimental section, a new experiment is carried out with a set of images extracted from the COLD dataset [72]. This publicly available dataset contains several sets of images that were captured while a mobile robot traversed a trajectory in some indoor environments. Therefore, the results in this section permit assessing the performance of the descriptors in a different environment and with a trajectory-like dataset.

To carry out the experiment, the Saarbrücken dataset is selected [72]. To create the training set, we have selected images from the Saarbrücken dataset in such a way that the distance between consecutive capture points is, on average, 30 cm. The rest of images are considered as test images, and they are used to solve the localization problem, as described in Section 3.

The results are presented in Figures 44 and 45. As in the previous experiments, we estimate both the position and the relative orientation of the robot and we consider either noise or occlusions in the test images. The descriptors included in this experiment are HOG, *gist*, WS and BG, since they have showed a good performance in the previous experiments. Additionally, their most relevant parameters are tuned with the values that provided, in general, best estimations in the previous subsections. The levels of noise or occlusion are the same than those included in the previous experiments: presence of different Gaussian noise ($\sigma^2 = \{0, 0.0025, 0.05, 0.01, 0.02\}$) and partial occlusions considering ($\{0, 5, 10, 20, 40\}$ %) of the image occluded.

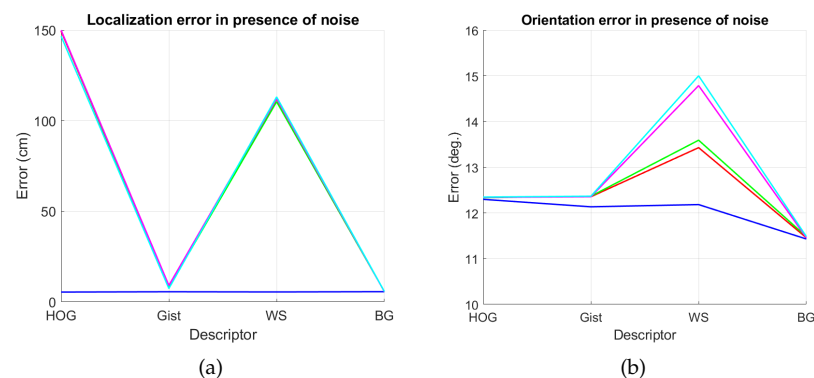


Figure 44. Average errors with the COLD dataset in the presence of noise. (a) Average position error (cm) and (b) average orientation error (deg). Legend: — Original, — Noise 1, — Noise 2, — Noise 3, — Noise 4.

First, Figure 44 shows (a) the average error of the localization task (expressed in cm) and (b) the average error of the orientation retrieval task (expressed in deg). Several levels of noise are considered in this experiment. Second, Figure 45 shows the same results but considering several levels of partial occlusions. It is worth highlighting that these errors cannot be directly compared with the absolute errors presented in the previous subsections, since the experimental setup is different. Notwithstanding that, these figures permit assessing the relative performance of the descriptors with a trajectory-like dataset and knowing if the descriptors present similar tendencies in different kinds of environments and datasets.

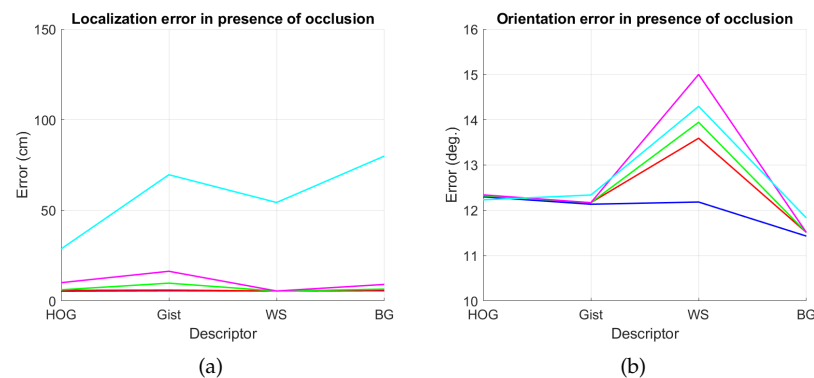


Figure 45. Average errors with the COLD dataset in the presence of occlusions. (a) Average position error (cm) and (b) average orientation error (deg). Legend : — Original, — Occlusion 1, — Occlusion 2, — Occlusion 3, — Occlusion 4.

Figure 44a shows that the relative performance of the descriptors when calculating the relative position in ideal conditions (i.e., with no noise) is quite similar. Additionally, *gist* and BG resist quite well the presence of noise. However, HOG and WS quickly degrade their performance as the level of noise increases. These results are in line with those presented in previous sections. About the relative orientation retrieval with noise, Figure 44b shows that HOG, *gist* and BG are quite robust, while WS performs worse with high levels of noise. Figure 45a proves that the four description methods present relatively good results in the presence of occlusions, except for the highest level of occlusion. In this case, HOG is the descriptor that best performs. About the orientation retrieval in presence of occlusions, Figure 45b shows that HOG, *gist* and BG perform well, independently on the level of occlusion, but WS quickly increases the error with high levels of occlusion.

6. Conclusions

This paper has focused on the study of the localization problem, using a previously built visual representation of the environment. The problem has been addressed as an absolute localization task, making use of the data provided by a catadioptric vision sensor mounted on the robot both to estimate both the position and the orientation of the robot. To extract relevant information from the images, methods based on the global appearance of the panoramic scenes have been implemented and assessed. A comparative evaluation has been carried out between six families of well-known global description methods.

The main contributions of the paper include an exhaustive study of global appearance techniques (FS, HOG, *gist*, WS, BG and RT) and the adaptation of some of these algorithms to store position and orientation information from panoramic scenes in such a way that both processes can be carried out sequentially. First, the position of the robot can be estimated and second, the orientation is estimated.

In addition, the computational cost to estimate the position and orientation has been studied, including the influence of the most relevant parameters. This study has revealed that FS and RT present a reasonable computational cost, and so do some specific configurations of HOG and *gist*, but *Wi-SURF* and *BRIEF-gist* are less competitive as far as computation time is concerned. From this point of view, FS, RT, HOG and *gist* could be feasible in real time applications. In addition to this, the performance of the descriptors has been tested in localization tasks. First, we have focused on the image retrieval problem. All the description methods have been tested along with several distance measures, and the results have shown that *Wi-SURF* and *BRIEF-gist* present the best relative results. Additionally, HOG with certain distance measures present very good results and the best relation between computational time and image retrieval rate. Second, the relative error of the position estimation has been studied. It has corroborated that: (a) HOG presents very good localization results under ideal conditions and is quite robust to noise and occlusions,

(b) Wi-SURF provides the most competitive results under ideal conditions but is very negatively influenced by noise and occlusions and (c) BRIEF-gist is very robust against these effects, but its results in ideal conditions are not remarkable. To finish, the problem of orientation estimation has been addressed. The best results are obtained with WS and BG but only when there is neither noise nor occlusions. If these phenomena are present, HOG and *gist* perform more robustly.

These results have demonstrated that global-appearance methods are a feasible approach to solve the localization task. Thanks to them, the robot can build a model of the environment and use it to estimate with accuracy the position and orientation of the robot in the environment, with computational efficiency. This fact may have interesting implications in future developments in the field of mobile robotics. As an example, this concept can be used to build hybrid maps that arrange the information in several layers, with different accuracy: a high level layer that permits carrying out a rough and quick localization and a lower layer that contains information with geometric accuracy and allows the robot to refine the estimation of its position. Global-appearance methods can be used on their own or in conjunction with feature-based techniques to develop algorithms that face these problems efficiently.

All these facts encourage us to go into this framework in depth. To build a fully autonomous mapping and localization system, several future works should be considered. First, the image collection process could be automated to obtain an optimal representation of the environment. Second, the mapping and localization processes could be integrated in a topological SLAM system that carries out both the model creation and the localization from the scratch. To optimize these algorithms we also consider carrying out a complete comparison between global-appearance and feature-based techniques as a future work.

Author Contributions: Conceptualization, L.P. and O.R.; methodology, L.P., O.R. and A.P.; software, L.P., M.B. and V.R.; validation, L.P., M.B. and V.R.; formal analysis, O.R. and A.P.; investigation, A.P., M.B. and V.R.; resources, L.P. and O.R.; data curation, M.B. and V.R.; writing—original draft preparation, L.P. and O.R.; writing—review and editing, V.R. and O.R.; visualization, A.P., M.B. and V.R.; supervision, L.P. and O.R.; project administration, L.P. and O.R.; funding acquisition, L.P. and O.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the Spanish Government through the project DPI2016-78361-R (AEI/FEDER, UE): *Creación de mapas mediante métodos de apariencia visual para la navegación de robots*, by the Generalitat Valenciana through the project AICO/2019/031: *Creación de modelos jerárquicos y localización robusta de robots móviles en entornos sociales* and by the Generalitat Valenciana and the FSE through the grant ACIF/2018/224.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <http://arvc.umh.es/db/images/quorumv/> (accessed on 9 May 2021).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Reinoso, O.; Payá, L. Special Issue on Mobile Robots Navigation. *Appl. Sci.* **2020**, *10*, 1317. [[CrossRef](#)]
2. Reinoso, O.; Payá, L. Special Issue on Visual Sensors. *Sensors* **2020**, *20*, 910. [[CrossRef](#)] [[PubMed](#)]
3. Junior, J.M.; Tommaselli, A.; Moraes, M. Calibration of a catadioptric omnidirectional vision system with conic mirror. *ISPRS J. Photogramm. Remote Sens.* **2016**, *113*, 97–105. [[CrossRef](#)]
4. Coors, B.; Paul Condurache, A.; Geiger, A. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 518–533.

5. Sun, C.; Hsiao, C.W.; Sun, M.; Chen, H.T. HorizonNet: Learning Room Layout With 1D Representation and Pano Stretch Data Augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
6. Pintore, G.; Agus, M.; Gobbetti, E. AtlantaNet: Inferring the 3D Indoor Layout from a Single 360 Image Beyond the Manhattan World Assumption. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 432–448.
7. Xu, S.; Chou, W.; Dong, H. A Robust Indoor Localization System Integrating Visual Localization Aided by CNN-Based Image Retrieval with Monte Carlo Localization. *Sensors* **2019**, *19*, 249. [[CrossRef](#)]
8. Leyva-Vallina, M.; Strisciuglio, N.; Lopez-Antequera, M.; Tylecek, R.; Blaich, M.; Petkov, N. TB-Places: A Data Set for Visual Place Recognition in Garden Environments. *IEEE Access* **2019**, *7*, 52277–52287. [[CrossRef](#)]
9. Cebollada, S.; Payá, L.; Flores, M.; Peidró, A.; Reinoso, O. A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data. *Expert Syst. Appl.* **2020**, *167*, 114195. [[CrossRef](#)]
10. Lowe, D. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
11. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [[CrossRef](#)]
12. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. BRIEF: Binary robust independent elementary features. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 778–792.
13. Leutenegger, S.; Chli, M.; Siegwart, R.Y. BRISK: Binary robust invariant scalable keypoints. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2548–2555.
14. Rublee, E.; Rabud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 ICCV 2011: International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
15. Alahi, A.; Ortiz, R.; Vanderghenst, P. FREAK: Fast Retina Keypoint. In Proceedings of the CVPR 2012: Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 510–517.
16. Yang, X.; Cheng, K.T.T. Local difference binary for ultrafast and distinctive feature description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 188–194. [[CrossRef](#)]
17. Krose, B.; Bunschoten, R.; Hagen, S.; Terwijn, B.; Vlassis, N. Visual homing in environments with anisotropic landmark distribution. *Auton. Robot.* **2007**, *23*, 231–245.
18. Menegatti, E.; Maeda, T.; Ishiguro, H. Image-based memory for robot navigation using properties of omnidirectional images. *Robot. Auton. Syst.* **2004**, *47*, 251–267. [[CrossRef](#)]
19. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [[CrossRef](#)]
20. Ulrich, I.; Nourbakhsh, I. Appearance-based place recognition for topological localization. In Proceedings of the IEEE International Conference on Robotics and Automation, San Francisco, CA, USA, 24–28 April 2000; pp. 1023–1029.
21. Amorós, F.; Payá, L.; Mayol-Cuevas, W.; Jiménez, L.M.; Reinoso, O. Holistic Descriptors of Omnidirectional Color Images and Their Performance in Estimation of Position and Orientation. *IEEE Access* **2020**, *8*, 81822–81848. [[CrossRef](#)]
22. Milford, M. Visual Route Recognition with a Handful of Bits. In Proceedings of the Robotics: Science and Systems, Sydney, NSW, Australia, 9–13 July 2012.
23. Berenguer, Y.; Payá, L.; Valiente, D.; Peidró, A.; Reinoso, O. Relative Altitude Estimation Using Omnidirectional Imaging and Holistic Descriptors. *Remote Sens.* **2019**, *11*, 323. [[CrossRef](#)]
24. Yuan, X.; Martínez-Ortega, J.F.; Fernández, J.A.S.; Eckert, M. AEKF-SLAM: A new algorithm for robotic underwater navigation. *Sensors* **2017**, *17*, 1174. [[CrossRef](#)] [[PubMed](#)]
25. Luthardt, S.; Willert, V.; Adamy, J. LLama-SLAM: Learning high-quality visual landmarks for long-term mapping and localization. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 2645–2652.
26. Cao, L.; Ling, J.; Xiao, X. Study on the Influence of Image Noise on Monocular Feature-Based Visual SLAM Based on FFDNet. *Sensors* **2020**, *20*, 4922. [[CrossRef](#)] [[PubMed](#)]
27. Shamsfakhr, F.; Bigham, B.S.; Mohammadi, A. Indoor mobile robot localization in dynamic and cluttered environments using artificial landmarks. *Eng. Comput.* **2019**, *36*, 400–419. [[CrossRef](#)]
28. Lin, J.; Peng, J.; Hu, Z.; Xie, X.; Peng, R. ORB-SLAM, IMU and Wheel Odometry Fusion for Indoor Mobile Robot Localization and Navigation. *Acad. J. Comput. Inf. Sci.* **2020**, *3*. [[CrossRef](#)]
29. Gil, A.; Mozos, O.M.; Ballesta, M.; Reinoso, O. A comparative evaluation of interest point detectors and local descriptors for visual SLAM. *Mach. Vis. Appl.* **2010**, *21*, 905–920. [[CrossRef](#)]
30. Dong, X.; Dong, X.; Dong, J.; Zhou, H. Monocular Visual-IMU Odometry: A Comparative Evaluation of Detector-Descriptor-Based Methods. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 2471–2484. [[CrossRef](#)]
31. Menegatti, E.; Zocaratto, M.; Pagello, E.; Ishiguro, H. Image-based Monte Carlo Localisation with Omnidirectional Images. *Robot. Auton. Syst.* **2004**, *48*, 17–30. [[CrossRef](#)]
32. Murillo, A.; Guerrero, J.; Sagües, C.; Filliat, D. Surf features for efficient robot localization with omnidirectional images. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Rome, Italy, 10–14 April 2007; pp. 3901–3907.

33. Siagian, C.; Itti, L. Biologically Inspired Mobile Robot Vision Localization. *IEEE Trans. Robot.* **2009**, *25*, 861–873. [[CrossRef](#)]
34. Payá, L.; Amorós, F.; Fernández, L.; Reinoso, O. Performance of Global-Appearance Descriptors in Map Building and Localization Using Omnidirectional Vision. *Sensors* **2014**, *14*, 3033–3064. [[CrossRef](#)]
35. Khaliq, A.; Ehsan, S.; Chen, Z.; Milford, M.; McDonald-Maier, K. A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes. *IEEE Trans. Robot.* **2019**, *36*, 561–569. [[CrossRef](#)]
36. Román, V.; Payá, L.; Cebollada, S.; Reinoso, Ó. Creating Incremental Models of Indoor Environments through Omnidirectional Imaging. *Appl. Sci.* **2020**, *10*, 6480. [[CrossRef](#)]
37. Marinho, L.B.; Almeida, J.S.; Souza, J.W.M.; Albuquerque, V.H.C.; Rebouças Filho, P.P. A novel mobile robot localization approach based on topological maps using classification with reject option in omnidirectional images. *Expert Syst. Appl.* **2017**, *72*, 1–17. [[CrossRef](#)]
38. Ma, J.; Zhao, J. Robust topological navigation via convolutional neural network feature and sharpness measure. *IEEE Access* **2017**, *5*, 20707–20715. [[CrossRef](#)]
39. Paya, L.; Reinoso, O.; Berenguer, Y.; Ubeda, D. Using omnidirectional vision to create a model of the environment: A comparative evaluation of global appearance descriptors. *J. Sens.* **2016**, *2016*, 1–21. [[CrossRef](#)]
40. Arroyo, R.; Alcantarilla, P.F.; Bergasa, L.M.; Yebes, J.J.; Bronte, S. Fast and effective visual place recognition using binary codes and disparity information. In Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 3089–3094.
41. Berenguer, Y.; Payá, L.; Peidró, A.; Gil, A.; Reinoso, O. Nearest Position Estimation Using Omnidirectional Images and Global Appearance Descriptors. In *Robot 2015: Second Iberian Robotics Conference*; Springer: Cham, Switzerland, 2016; pp. 517–529.
42. Ishiguro, H.; Tsuji, S. Image-based memory of environment. In Proceedings of the 1996 IEEE/RSJ International Conference on Intelligent Robots and Systems '96 (IROS 96), Osaka, Japan, 4–8 November 1996; Volume 2, pp. 634–639. [[CrossRef](#)]
43. Sturzl, W.; Mallot, H. Efficient visual homing based on Fourier transformed panoramic images. *Robot. Auton. Syst.* **2006**, *54*, 300–313. [[CrossRef](#)]
44. Horst, M.; Möller, R. Visual place recognition for autonomous mobile robots. *Robotics* **2017**, *6*, 9. [[CrossRef](#)]
45. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume II, pp. 886–893.
46. Zhu, Q.; Avidan, S.; Yeh, M.C.; Cheng, K.T. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1491–1498. [[CrossRef](#)]
47. Hofmeister, M.; Liebsch, M.; Zell, A. Visual self-localization for small mobile robots with weighted gradient orientation histograms. In Proceedings of the 40th International Symposium on Robotics, Barcelona, Spain, 10–13 March 2009; IFR: Frankfurt am Main, Germany, 2009; pp. 87–91.
48. Hofmeister, M.; Vorst, P.; Zell, A. A comparison of Efficient Global Image Features for Localizing Small Mobile Robots. In Proceedings of the 41st International Symposium on Robotics, Munich, Germany, 7–9 June 2010; pp. 143–150.
49. Aslan, M.F.; Durdu, A.; Sabanci, K.; Mutluer, M.A. CNN and HOG based comparison study for complete occlusion handling in human tracking. *Measurement* **2020**, *158*, 107704. [[CrossRef](#)]
50. Neumann, D.; Langner, T.; Ulbrich, F.; Spitta, D.; Goehring, D. Online vehicle detection using Haar-like, LBP and HOG feature based image classifiers with stereo vision preselection. In Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 11–14 June 2017; pp. 773–778.
51. Payá, L.; Fernández, L.; Gil, A.; Reinoso, O. Map Building and Monte Carlo Localization Using Global Appearance of Omnidirectional Images. *Sensors* **2010**, *10*, 11468–11497. [[CrossRef](#)]
52. Oliva, A.; Torralba, A. Building the gist of scenes: The role of global image features in recognition. *Prog. Brain Res. Spec. Issue Vis. Percept.* **2006**, *155*, 23–36.
53. Torralba, A. Contextual priming for object detection. *Int. J. Comput. Vis.* **2003**, *53*, 169–191. [[CrossRef](#)]
54. Siagian, C.; Itti, L. Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 300–312. [[CrossRef](#)]
55. Chang, C.K.; Siagian, C.; Itti, L. Mobile robot vision navigation and localization using Gist and Saliency. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010; pp. 4147–4154. [[CrossRef](#)]
56. Murillo, A.; Singh, G.; Kosecka, J.; Guerrero, J. Localization in Urban Environments Using a Panoramic Gist Descriptor. *IEEE Trans. Robot.* **2013**, *29*, 146–160. [[CrossRef](#)]
57. Liu, Y.; Zhang, H. Visual loop closure detection with a compact image descriptor. In Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vilamoura-Algarve, Portugal, 7–12 October 2012; pp. 1051–1056.
58. Su, Z.; Zhou, X.; Cheng, T.; Zhang, H.; Xu, B.; Chen, W. Global localization of a mobile robot using lidar and visual features. In Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), Macau, Macao, 5–8 December 2017; pp. 2377–2383.

59. Andreasson, H.; Treptow, A.; Duckett, T. Localization for mobile robots using panoramic vision, local features and particle filter. In Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 3348–3353.
60. Agrawal, M.; Konolige, K.; Blas, M.R. Censure: Center surround extremas for realtime feature detection and matching. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 102–115.
61. Badino, H.; Huber, D.; Kanade, T. Real-time topometric localization. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation, Saint Paul, MN, USA, 14–18 May 2012; pp. 1635–1642.
62. Zhang, M.; Han, S.; Wang, S.; Liu, X.; Hu, M.; Zhao, J. Stereo Visual Inertial Mapping Algorithm for Autonomous Mobile Robot. In Proceedings of the 2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE), Oxford, UK, 10–12 August 2020; pp. 97–104.
63. Aladem, M.; Rawashdeh, S.A. Lightweight visual odometry for autonomous mobile robots. *Sensors* **2018**, *18*, 2837.
64. Sünderhauf, N.; Protzel, P. Brief-gist-closing the loop by simple means. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 1234–1241.
65. Radon, J. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Class. Pap. Mod. Diagn. Radiol.* **2005**, *5*, 21.
66. Hoang, T.V.; Tabbone, S. A geometric invariant shape descriptor based on the Radon, Fourier, and Mellin transforms. In Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 23–26 August 2010; pp. 2085–2088.
67. Hasegawa, M.; Tabbone, S. A shape descriptor combining logarithmic-scale histogram of radon transform and phase-only correlation function. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 182–186.
68. Berenguer, Y.; Payá, L.; Ballesta, M.; Reinoso, O. Position estimation and local mapping using omnidirectional images and global appearance descriptors. *Sensors* **2015**, *15*, 26368–26395. [[PubMed](#)]
69. Juliá, M.; Gil, A.; Reinoso, O. A comparison of path planning strategies for autonomous exploration and mapping of unknown environments. *Auton. Robot.* **2012**, *33*, 427–444. [[CrossRef](#)]
70. Liu, S.; Li, S.; Pang, L.; Hu, J.; Chen, H.; Zhang, X. Autonomous Exploration and Map Construction of a Mobile Robot Based on the TGHM Algorithm. *Sensors* **2020**, *20*, 490. [[CrossRef](#)] [[PubMed](#)]
71. ARVC. Automation, Robotics and Computer Vision Research Group. Miguel Hernández University. Spain. Quorum 5 Set of Images. Available online: <http://arvc.umh.es/db/images/quorumv/> (accessed on 29 December 2020).
72. Pronobis, A.; Caputo, B. COLD: COsy Localization Database. *Int. J. Robot. Res. (IJRR)* **2009**, *28*, 588–594. [[CrossRef](#)]