IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 2, 2020, accepted April 23, 2020, date of publication April 28, 2020, date of current version May 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990996

Holistic Descriptors of Omnidirectional Color Images and Their Performance in Estimation of Position and Orientation

FRANCISCO AMORÓS[1], LUIS PAYÁ[1], WALTERIO MAYOL-CUEVAS[2], (Member, IEEE), LUIS MIGUEL JIMÉNEZ[1], AND OSCAR REINOSO[1], (Senior Member, IEEE)

[1]Department of Systems Engineering and Automation, Miguel Hernández University, 03202 Elche, Spain
[2]Department of Computer Science, University of Bristol, Bristol BS81TH, U.K.

Corresponding author: Luis Payá (lpaya@umh.es)

This work was supported in part by the Spanish Government through the Project DPI 2016-78361-R (AEI/FEDER, UE) "Creación de mapas mediante métodos de apariencia visual para la navegación de robots", and in part by the Generalitat Valenciana through the Project AICO/2019/031 "Creación de modelos jerárquicos y localización robusta de robots móviles en entornos sociales".

ABSTRACT The use of visual sensors in robotic navigation tasks is a common approach, and numerous examples can be found in the literature. This work focuses on the problem of map building and localization using omnidirectional images as the only source of information. The main objective of this paper is to present a thorough comparison of global-appearance description techniques including the use of color information in different approaches. Some of the descriptors have been widely tested in previous works using gray-level images. In the present work we concentrate on the role and efficiency of the color information. Other descriptors are presented for the first time. To carry out this study, a database captured in different areas of an office environment is used, including two different datasets: training and test datasets. The experimental results include computational requirements in the map building and localization processes, and the accuracy in the pose estimation of the test images in a topological map, separating both position and orientation. To complete the study, the behavior of the descriptors is tested when the images present noise or occlusions, specially the effect on the color information.

INDEX TERMS Catadioptric vision sensors, global-appearance descriptors, color images, Fourier signature, topological mapping, histogram of oriented gradients, *gist*.

I. INTRODUCTION

The autonomous navigation of mobile robots is a wide area of investigation. For this task, robots must gather and interpret information from their environment. In the literature, different approximations can be found depending on the kind of sensors used. Over the last few years, an important line of research is the use of visual sensors [1], due to the many possibilities they offer, the richness of the information they provide, and their suitability for this purpose, since they consume less power than other sensors, which is important for the autonomy of navigation, and their cost is relatively low.

Visual systems can be classified depending on the number of cameras they use and their field of view. That way, we find examples of systems based on one camera [2], [3], stereo cameras [4], [5] that simulate the human vision, trinocular systems [6] or even arrays of cameras that gather the 90% of the spherical field of view around the robot [7]. If we consider also adaptations in the architecture of the visual sensors, catadioptric systems can be highlighted [8]. They use a reflective surface to expand the visual field of view [9], [10].

The richness of the visual information implies important memory and computational requirements to store and process the scenes. In real-time navigation tasks, this quantity of information might become unmanageable. For that reason, it is necessary to represent the images using descriptors that reduce the information to a vector of features, but preserve the ability to recognize the image among others in a database.

Such descriptors can be classified into two categories: local-features descriptors and global-features or holistic descriptors. On the other hand, local-features descriptors extract outstanding points from the images, which the robot

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

can recognize easily. These features are also called landmarks. Landmarks can be artificial, as Okuyama *et al.* show in [11] using QR codes, or natural. Natural landmarks are extracted directly from the image, and usually correspond to recognizable points as corners, doors or windows, as we can see in [12], [13]. Another example is found in [14], where a novel method for object recognition and pose estimation based on 3D point extraction using an RGB-D sensor is presented. The main disadvantage of these techniques is the complexity in the extraction of stable landmarks in real and changing environments, and the computational cost of processing the image to extract those features and comparing them.

On the other hand, global-appearance descriptors extract the information of the image as a whole, avoiding any local pattern of the scene. Map building and localization with these descriptors is less complex than using 3D landmarks [15]. However, the size of the maps can be excessive, since they contain information of the entire image. That way, the study of global-appearance descriptors normally focuses on the kind and quantity of features they extract from the images. In contrast with the descriptors based on landmarks, they do not contain any metric information. For that reason, they are typically used for topological navigation approaches, in which the localization of the robot can be addressed as an image association problem with the information in the map [16]. Several authors have addressed a variety of problems in autonomous vehicles using visual information and global-appearance descriptors. For example, Hu *et al.* [17] use holistic descriptors from images, with the purpose of recognizing signals in road environments. They build these holistic descriptors from local features and a method based on the k-nearest neighbours. Payá *et al.* [18] present a framework for topologic map creation using global-appearance descriptors. Additionally, these description techniques can be combined with clustering algorithms in order to improve the maps and the localization process, as [19] shows.

Image retrieval plays an important role in robot localization, and this problem has been extensively addressed using grayscale images and holistic descriptors. Li *et al.* [20] study the image matching problem, with the objective of detecting loop closures in a SLAM (Simultaneous Localization and Mapping) application. They solve it by using a combination of clustering methods and descriptors built both from holistic and local features of grayscale images. Horst and Möller [21] focus on place recognition in mobile robotics using grayscale images. They investigate the effect of warping in place recognition and the NSAD (Normalized Sum of Absolute Differences) distance measure. Doan *et al.* [22] also study the problem of place recognition using visual information and exploiting the temporal continuity of the acquisition process. The image retrieval pipeline uses local features and an encoding method that represents each image as a single vector.

A recent approximation showed in [23] demonstrates that it is possible to estimate relative positions between images using global-appearance techniques. The framework, called multi-scale analysis, uses plane projections of the omnidirectional images that permit estimating displacements between two positions of the robot using only visual information. That way, it improves the accuracy of the robot's localization in the map.

The descriptors included in this work are based on Discrete Fourier Transform [24], the Histogram of Oriented Gradients [25] and *gist* [26]. Most of the descriptors that can be found in the literature are designed to be used with greyscale images. In the present work we explore the role of color information along with global-appearance descriptors [27] and we assess the performance of such information in a topological localization task, addressed as an image retrieval problem.

The remainder of the article is structured as follows: section II introduces the global-appearance techniques used to describe the omnidirectional images in this work. Section III outlines the introduction of color features to the description techniques. Later, section IV presents the sets of images used in the experiments. Section V details the experimental setup. After that, Section VI presents the results, and finally, section VII summarizes the main conclusions.

II. VISUAL DESCRIPTORS

This section introduces the techniques used to describe globally the appearance of the panoramic images in the present work. Some of them have been extensively described in previous works: the Fourier Signature (FS) and the Histogram of Oriented Gradients (HOG) in [28] and Principal Components Analysis (PCA) and *gist* in [29]. In this work, this set of techniques is complemented with two additional descriptors based on the Discrete Fourier Transform and one additional *gist* descriptor based on color information [30]. These techniques are outlined in the next subsections. In all the cases, the initial information is a set of N panoramic images captured from several points in the ground plane, distributed along the environment to model $\mathfrak{F} = \{f_1, f_2, \ldots, f_N\}$, where $f_j \in \mathbb{R}^{N_x \times N_y}, j = 1, \ldots, N$ represent each image of the map set. N_x and N_y denote, respectively, the number of rows and columns of the image f_j. In general, after describing each of these images, the result is a set of position descriptors, one per original scene $\mathfrak{D}^{pos} = \{\vec{d}_1^{pos}, \vec{d}_2^{pos}, \ldots, \vec{d}_N^{pos}\}$, where $\vec{d}_j^{pos} \in \mathbb{R}^{k^{pos} \times 1}$ and a set of orientation descriptors, also one per original scene $\mathfrak{D}^{or} = \{\vec{d}_1^{or}, \vec{d}_2^{or}, \ldots, \vec{d}_N^{or}\}$, where $\vec{d}_j^{or} \in \mathbb{R}^{k^{or} \times 1}$. k^{pos} is the size of the position descriptor and k^{or} is the size of the orientation descriptor. Their specific values depend on each description technique, as described in Section V-B.

A. TECHNIQUES BASED ON THE DISCRETE FOURIER TRANSFORM

The Discrete Fourier Transform (DFT) converts the sequence of numbers $\{a_0, a_1, \ldots, a_{N_y-1}\}$ in the complex sequence $\{A_0, A_1, \ldots, A_{N_y-1}\}$ according the equation:

$$A_k = \sum_{n=0}^{N_y-1} a_n \cdot e^{-j\frac{2\pi}{N_y}kn}; \quad k = 0, \ldots, N_y - 1, \qquad (1)$$

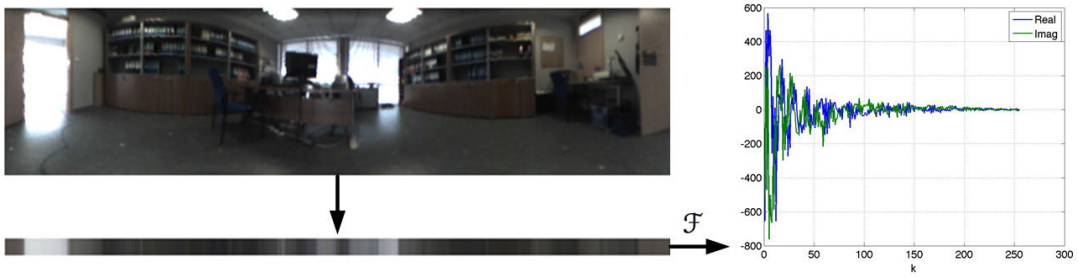


FIGURE 1. Process to obtain the 1D-DFT descriptor of a panoramic image.

where N_y is the number of components of the sequence. This transformation represents a discrete signal in the frequency domain. One relevant property for this work is the shift theorem, which states that a circular shift of the initial sequence produces a transformed sequence whose components have the same magnitude and the arguments can be calculated with (2).

$$\mathcal{F}[\{a_{n-q}\}]_k = A_k e^{-j\frac{2\pi qk}{N_y}}; \quad k = 0, \ldots N_y - 1 \tag{2}$$

where q is the amount of circular shift in the first sequence.

1) ONE-DIMENSIONAL DFT (1D-DFT)

Briggs *et al.* [31]–[33] propose a descriptor that reduces a panoramic image into a unidimensional vector for localization and navigation purposes in robotics. From these works, we develop the idea of creating a one-dimensional vector from the average values of the pixels of each column of the panoramic image. After that, we apply the DFT to the resulting vector. Fig. 1 shows the descriptor creation process. This process is applied individually to each panoramic image in the initial set $\mathfrak{F}$.

If the movement of the robot is contained in the ground plane and the catadioptric vision system is mounted vertically, the 1D-DFT descriptor presents interesting properties when it is applied to the panoramic images obtained from this system. First, the most relevant information of the image is contained in the lowest frequency components so only a number of components is usually retained. Moreover, the last components are usually affected by the presence of high-frequency noise in the original image. For that reason, we keep only the first components, having a substantial compression effect. Second, since the transformed sequence is complex, the information can be separated into two vectors: one with the magnitudes and the other with the arguments.

Additionally, according to the shift theorem of the DFT in (2), the magnitudes vector is invariant against changes of the orientation of the robot in the ground plane and can be used for localization purposes, while the arguments vector retains information of phase that is useful in the estimation of the relative orientation of the robot. In this theorem, if the robot rotates θ degrees, the sequence that represents the original panoramic image circularly shifts q positions. Therefore, the magnitudes vector can be considered as the position descriptor (it contains information on the appearance of the environment as seen from a specific position, independently on the orientation), and the arguments vector can be considered as the orientation descriptor (it is useful to estimate the relative orientation of the robot with respect to a reference one).

2) FOURIER SIGNATURE (FS)

Ishiguro and Tsuji [34] proposed the creation of visual maps using the DFT of each row of a panoramic image. This descriptor is also used in [24] with the name of Fourier Signature (FS). The FS is a complex matrix and it is also calculated independently for each panoramic image in the initial set $\mathfrak{F}$. Using the same property than in the previous subsection, from each row, only the first terms of the transform are retained. The resulting magnitudes and arguments matrices are arranged into two vectors to compose, respectively, the position and the orientation descriptor of each initial panoramic image.

3) TWO-DIMENSIONAL DFT (2D-DFT)

Finally, it is also possible to apply the 2D-DFT directly over a digital image to transform the visual information into the frequency domain. If we represent an image with the discrete function $f(x, y)$, with N_x rows and N_y columns, the 2D-DTF is obtained as:

$$\begin{aligned}\mathcal{F}[\{f(x,y)\}] = F(u,v) = \frac{1}{N_y N_x}\sum_{x=0}^{N_x-1}\sum_{y=0}^{N_y-1} f(x,y) \\ \cdot e^{-j2\pi\left(\frac{ux}{N_x}+\frac{vy}{N_y}\right)} \\ u = 0, \ldots N_x - 1, \quad v = 0, \ldots N_y - 1.\end{aligned} \tag{3}$$

Like in the previous DFT-based descriptor, the coefficients of the transform can be divided in two matrices, one with the magnitudes (or power spectrum) which is useful as position descriptor, and other with the arguments, which is the orientation descriptor. A pure rotation of the robot in the floor plane produces a shift of the columns of the panoramic images. The shift theorem of the 2D-DFT is expressed as:

$$\begin{aligned}\mathcal{F}[\{f(x-x_0, y-y_0)\}] = F(u,v)\cdot e^{-j2\pi\left(\frac{ux}{N_x}+\frac{vy}{N_y}\right)} \\ u = 0, \ldots N_x - 1, \quad v = 0, \ldots N_y - 1.\end{aligned} \tag{4}$$

In this case, to compose the final descriptor, a number of low-frequency components is retained (i.e. a submatrix starting from the first component of the transform).

B. TECHNIQUES BASED ON PCA

Principal Component Analysis (PCA) is a technique which is widely used to extract the most relevant information from a set of data vectors, which consists in performing a transformation that projects these data vectors into a lower-dimensional space that preserves most of the variance of the data [35].

The pixels of an image can be arranged into a column vector $\vec{x} \in \mathbb{R}^{M\times 1}$, with M the number of elements of the image. Considering N the number of images of the dataset, the matrix of data is denoted as $X = [\vec{x}_1|\vec{x}_2|\ldots|\vec{x}_N] \in \mathbb{R}^{M\times N}$. To perform PCA, we normalize the data by subtracting the average value from each image. We denote the new matrix as $\widehat{X}$. From these data, the covariance matrix is obtained $C = \frac{1}{N}\,\widehat{X}\cdot\widehat{X}^T$ with $C \in \mathbb{R}^{M\times M}$. From the eigenvectors of this matrix $\vec{u}_j$, ordered by the relative importance of their associated eigenvalues, we obtain the transformation matrix:

$$U = [\,\vec{u}_1 \quad \vec{u}_2 \quad \ldots \quad \vec{u}_N\,], \tag{5}$$

with $U \in \mathbb{R}^{M\times N}$. The projection of the original information in the new basis is:

$$\widehat{Y} = U^T \cdot \widehat{X} \tag{6}$$

where

$$\widehat{Y} = [\vec{\hat{y}}_1 \quad \vec{\hat{y}}_2 \quad \ldots \quad \vec{\hat{y}}_N] \tag{7}$$

$\vec{\hat{y}}_j \in \mathbb{R}^{N\times 1}$ is the projection of $\vec{\hat{x}}_j$ in the new basis. In practice, we select only the first eigenvectors $\vec{u}_j$ to build the new basis.

1) ROTATIONAL PCA

PCA has demonstrated to be a robust algorithm in the compression of information. However, if PCA was used directly, considering that the data vectors are panoramic images captured from different positions, then the projections would not be rotationally invariant. That is to say, the projection of two images captured from the same position but with different robot orientations would lead to completely different projections in the new space.

To solve this problem, Jogan and Leonardis [36], [37] propose the *Eigenspace of Spining-Images*. The algorithm makes R_{im} equally distributed artificial rotations of each original panoramic scene and builds the initial data matrix with them. Using the algorithm they propose, every image is transformed into a column vector (also named 'projection') whose components are complex numbers. Figure 2 represents that, in the case of a panoramic image and its evenly rotated siblings, every specific component of their projections has the same magnitude, and these components have a phase lag which is constant between consecutive rotated siblings. More concisely, the blue asterisks show the second component of the projection of a panoramic image and the second component of the projections of its rotated siblings. The magnitude

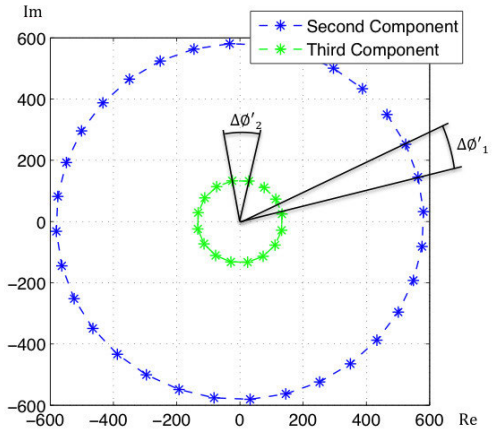


FIGURE 2. Second and third components of the projections of an image and its 31 rotated siblings.

of these second components is the same, and there is a phase lag between the second component of the projection of consecutive rotated siblings which is constant and equal to $\Delta\phi'_1$. The green asterisks show the same concept by representing the third component of the projection of a panoramic image and their rotated siblings. Again, there is a phase lag between the third component of the projection of consecutive rotated siblings which is constant and equal to $\Delta\phi'_2$. Therefore, the map only needs to contain the projection of one image per position (position descriptors), and the phase lag between the coefficients of consecutive rotated images (orientation descriptors). That way, we can artificially simulate the projections of the different rotations of the image. The magnitude will be used to find the location of the robot in the map, and the argument information to estimate the orientation. Additionally, it is necessary to store the transformation matrix U.

The angular resolution of the dataset will depend on the number of artificial rotations included in the map, according to (8). However, high resolutions will require an extremely high calculation time to obtain the projections.

$$Min.\ Angle(^\circ) = \frac{360}{R_{im}} \tag{8}$$

2) PCA OVER THE FOURIER SIGNATURE

As stated before, PCA is a technique which is not rotationally invariant. However, if the information of the data matrix X presents rotational invariance, the new representation will also keep this property, as stated in [38], [39]. For this reason, in this section we propose the next method. For each original panoramic image, we calculate the magnitudes matrix of its FS and we arrange the information in a column vector. The data matrix X will be composed of the column vectors obtained from the set of panoramic images, and PCA is subsequently performed with this matrix. The projections in the new space will be used for the robot localization. For the orientation estimation, the descriptor uses the arguments of the Fourier Signature, without any change of basis.

C. TECHNIQUES BASED ON HISTOGRAMS OF ORIENTED GRADIENTS

The Histogram of Oriented Gradients (HOG) [40] describes the image using the pixel intensity distribution in local areas. For that purpose, first, the gradient of the image is obtained. If $\mathcal{I}_x$ and $\mathcal{I}_y$ represent the derivatives of the image regarding axis x and y respectively, it is possible to calculate the magnitude and orientation of the gradient as:

$$|G| = \sqrt{\mathcal{I}_x^2 + \mathcal{I}_y^2} \tag{9}$$

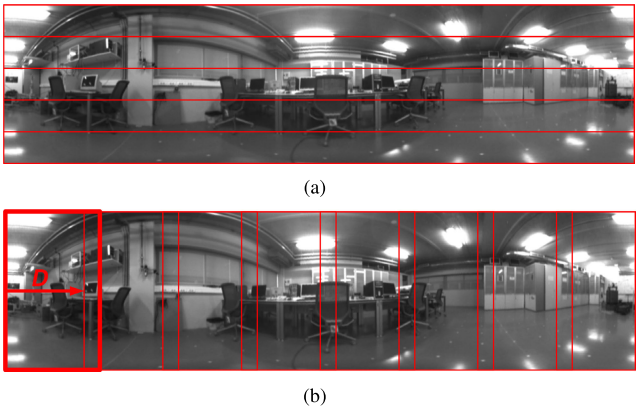
$$\theta = atan\frac{\mathcal{I}_y}{\mathcal{I}_x} \tag{10}$$

After that, the image is divided in cells and an histogram of oriented gradient per cell is compiled. The histogram of each cell is built from the information of the gradient orientation of each pixel in the cell, weighted by the gradient magnitude of this pixel. To build the histogram, a number of bins must be defined. In this work, we divide the orientation range (0° to 180°) into 8 bins, i.e. each 22.5°.

In order to adapt the technique to localization and orientation estimation purposes, we create two different descriptors: one for position and another for orientation estimation. Since we work with panoramic images, which contain the same information per row independently on the robot orientation, we use horizontal cells (with the same width than the image) to obtain a position descriptor, which presents rotational invariance. Regarding the orientation, we use overlapped vertical cells (with the same height than the image), separated a distance of D pixels between consecutive cells. By shifting the histograms of these cells, we can simulate a rotation of the robot. The resolution in the phase estimation depends on D:

$$\text{Min. Angle}(^\circ) = \frac{D \cdot 360}{N_y} \tag{11}$$

Fig. 3 shows the division of the image in cells, both for the position and the orientation descriptors. The descriptor will contain the histograms of each cell, appended and arranged in a column vector.



(a) (b)

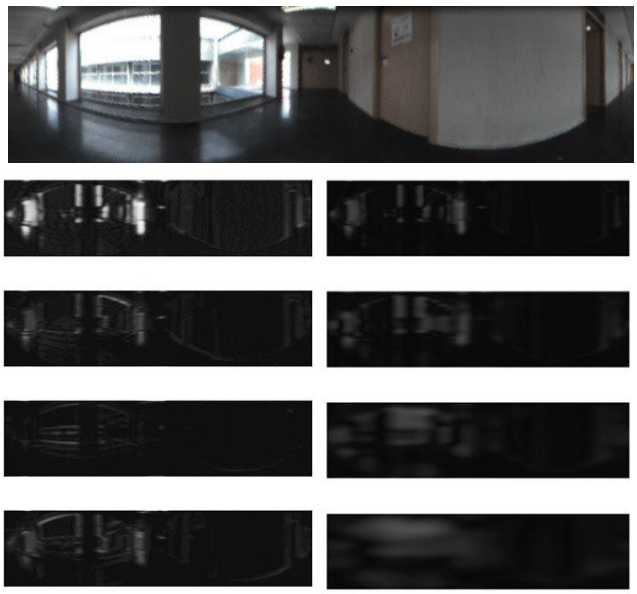
FIGURE 3. Cell divisions of a panoramic image to obtain the (a) position and (b) orientation descriptor.

D. TECHNIQUES BASED ON GIST

To obtain the essential information from the image, the descriptors based on *gist* try to mimic the human perception system and its ability to recognize a scene through the identification of colour or remarkable structures, avoiding the representation of specific objects or local features ([41], [42]). Therefore, they can be seen as global-appearance descriptors. In this work, we consider two approaches: *gist*-Gabor and *gist*-color.

1) GIST-GABOR

The *gist*-Gabor descriptor [26] is based on the use of Gabor filters, and collects frequency and orientation information from the images. The first step is to create a bank of Gabor filters, with orientations evenly distributed in the range [0°, 180°[. Gabor masks are frequency waves multiplied by a Gaussian function, so they are determined both in frequency and space domain. In this work, two different spatial scales are considered to create the Gabor bank. Fig. 4 presents a sample panoramic image and the resulting images after filtering it with four different Gabor masks, changing both scales and orientations.



(a) (b)

FIGURE 4. Sample panoramic image and resulting images after filtering it with four Gabor masks with different orientations (0°, 45°, 90°, 135°) and spatial scales: (a) scale 1 and (b) scale 2.

Once the scene has been filtered with the different masks and scales, the algorithm divides each resulting image into a set of (a) non-overlapping horizontal blocks, to create the position descriptor and (b) overlapping vertical blocks, to create the orientation descriptor, as seen in Fig. 3, and the average value of the pixels inside each block is calculated. Like in the case of the HOG descriptor, the resolution of our descriptor in orientation estimation depends on the distance between consecutive vertical blocks D, as seen in (11).

2) GIST-COLOR

The second descriptor based on *gist* is *gist*-color [43]. This technique collects color, intensity and orientation information from each scene. The color features are extracted from a Gaussian pyramid of images, using the color channels proposed by Hering [44], that defines three opposing color pairs: red/green, blue/yellow and black/white. The last one corresponds to the intensity of the pixel. The descriptor calculates five primary channels: R (Red), G (Green), B (Blue), Y (Yellow) and I (Intensity).

$$R = r - \frac{(g+b)}{2} \tag{12}$$

$$G = g - \frac{(r+b)}{2} \tag{13}$$

$$B = b - \frac{(r+g)}{2} \tag{14}$$

$$Y = r + g - 2 \cdot (|r-g| + b) \tag{15}$$

$$I = \frac{(r+g+b)}{3} \tag{16}$$

where r, g, b are the red, green and blue channels in the original RGB panoramic scene. The opposing color pairs are obtained from the primary colors as:

$$RG = |R - G| \tag{17}$$

$$BY = |B - Y| \tag{18}$$

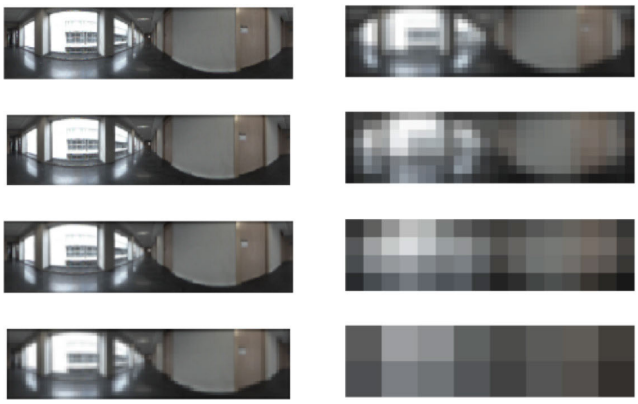


FIGURE 5. Gaussian Pyramid of a sample panoramic image with 8 scales, defined to carry out center-surround operations.

After that, a Gaussian pyramid is used to carry out a set of center-surround operations with the three opposing color channels RG, BY and I (21). Fig. 5 shows a Gaussian pyramid with 8 scales, created from a sample panoramic image. In these operations, the *center* corresponds to the lower scales (with higher resolution), that is denoted by c in (21). For the *surrounding* pixels (s), the lower resolution scales are used. The comparison between scales is represented with $\ominus$:

$$RG(c,s) = |(R(c) - G(c)) \ominus (R(s) - G(s))| \tag{19}$$

$$BY(c,s) = |(B(c) - Y(c)) \ominus (B(s) - Y(s))| \tag{20}$$

$$I(c,s) = |I(c) \ominus I(s)| \tag{21}$$

Using the center-surround operations, we obtain information in different scales which is expected to be robust against changes of lighting conditions, as Siagian *et al.* state in [45]. The scales used in the center-surround operations in this work are summarized in Table 1. Fig. 6 shows the resulting images after applying the center-surround operations with the three opposing color pairs of a sample image.

TABLE 1. Scales of the Gaussian pyramid that are used to carry out the center-surround operations.

c	s
2	3
2	4
2	5
2	6
3	4
3	6

The features of spatial distribution of the scenes, they are extracted using Gabor filters. For *gist*-color, we use 4 filter orientations ($\theta_i = 0°, 45°, 90°, 135°$) applied to two different pyramid scales. Finally, all the resulting images (both those with the color and those with the orientation information) are individually blockified. Like in the previous subsection (*gist*-Gabor), two descriptors are created: one with the values of the horizontal cells for localization purposes, and another with the values of the vertical cells for orientation estimation.

III. GLOBAL-APPEARANCE DESCRIPTORS AND COLOR INFORMATION

The descriptors included in Section II, with the exception of *gist*-color, extract the information from the scenes using only the gray-level intensity of each pixel. In fact, the great majority of global-appearance descriptors in the bibliography are applied only to grayscale images. However, if the images are captured with a color camera, the information provided by the different color channels can be used with the aim of improving the descriptors with more insightful information from the scene.

Initially, we can take advantage of the color information by applying the same description method separately to each of the three RGB channels. However, there is usually a high correlation among the information of these three channels. As a result, it is expected that the different descriptors also present a high correlation between them. If that happens, this would not add any useful information with respect to grayscale. As an example, Fig. 7(a) shows the values of the HOG descriptor applied to the same image in grayscale, and applied to the R, G and B channels of the same scene separately. As shown, a high correlation exists between the four descriptors. In this case, creating the descriptor of each RGB channel is almost equivalent to repeating three times the information of the grayscale descriptor. Additionally, Fig. 7(b) presents the same comparison but using the HSV channels (Hue, Saturation and Value). As expected, the descriptor of

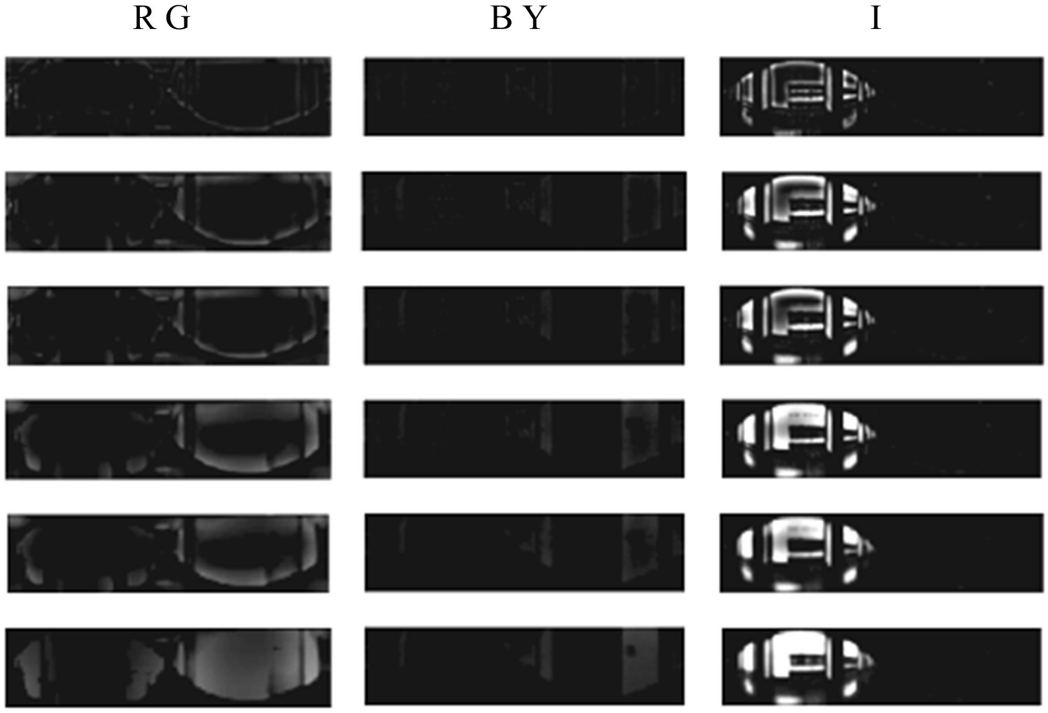


FIGURE 6. Center-surround operations using different scales for RG, BY and I channels of a sample image.

channel V is the same that grayscale space. However, H and S provide different information.

For this reason, we suggest other means of using the color information in order to extract useful features. In the literature, we can find several works that use the HSV color space. For example, Sablak and Bould [46] create a descriptor with the histograms of the image values in HSV space. Specifically, the descriptor is made up of the position of the local maxima of the histograms of channels H, S and V of the image separately. Suhasini *et al.* [47] also use HSV instead of RGB in order to obtain a descriptor based on the combination of SIFT (Scale Invariant Feature Transform) and ICH (Invariant Color Histogram), presenting an important improvement in image association tasks compared with the same algorithm applied to RGB. Junhua and Jing [48] show an image classification algorithm based on the Contourlet Transform using the H channel in the HSV space.

The color information of the scene can also be represented with the values of the pixels of each channel using histograms. These features are also independent on the scale and resolution of the image. With the aim of creating a useful descriptor, we propose to extract features by dividing the image into cells and building a histogram per cell using the information in the color channels. For localization, we divide the image in horizontal cells, as we do to obtain the HOG and *gist* descriptors (Fig. 3). This way, the resulting color descriptors are rotationally invariant since, from a specific position of the robot in the environment, they contain the same information, independently on the robot orientation.

Therefore, for each cell and channel of the color scene, a new histogram with the pixel intensity values is created. All these histograms are put together to create the final descriptor. We name this descriptor Color Histogram (CH). The bins that divide the histogram are equally distributed along the range of values of each channel. We also normalize the histograms by dividing the values of the bins by the number of pixels included in the cell. The size of the CH descriptor will directly depend on the number of cells of each image, and the bins of each histogram.

We can append the CH information to the descriptors that result from each of the techniques presented in Section II, obtaining complete descriptors that contain information both about the spatial distribution of the scene and about color. Specifically, we build a descriptor per scene using either a DFT-based method, HOG or *gist*, as presented in Section II and subsequently append the CH information. Before appending the color information to compose the final descriptor, we normalize each vector separately. This way, we avoid that any of the two parts weights excessively due to the number of components or the different magnitudes of each part. Regarding the normalization of color information, we take it into account both the number of histograms included in the descriptor CH, and the number of cells into which the image is divided.

We define $\vec{h}_j^H$, $\vec{h}_j^S$ and $\vec{h}_j^V$ as the column vectors that contain the values of the histograms of the channels H, S and V, respectively, compiled in the cell j. Each histogram is divided by the number of pixels of the cell. Then, we define h_{color_j} as

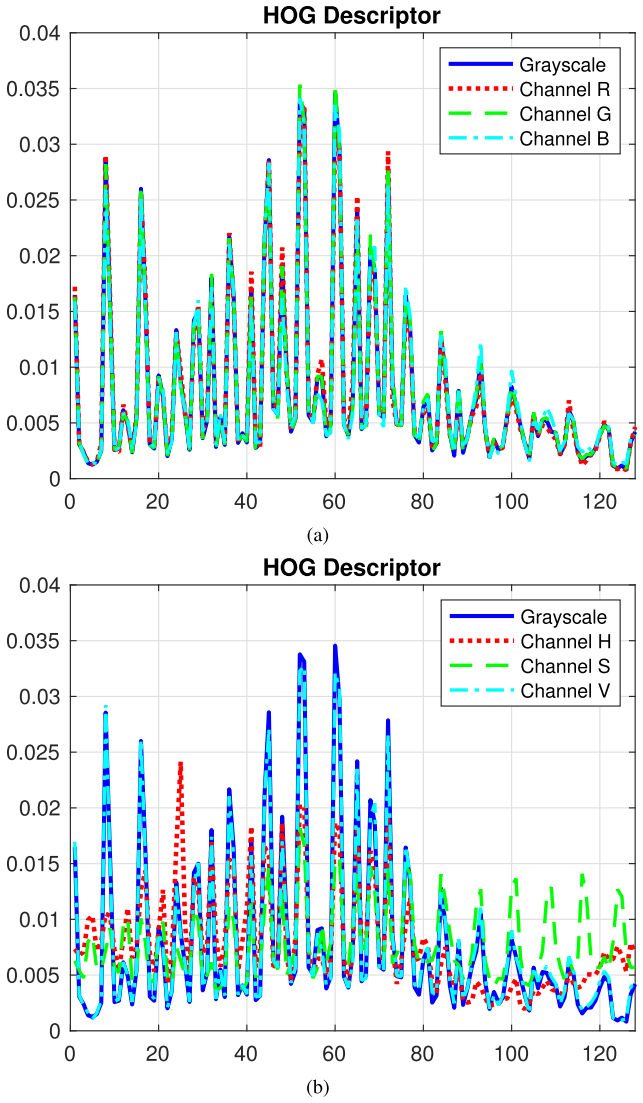


FIGURE 7. HOG descriptor of a sample image. The horizontal axis shows the number of component in the descriptor. Comparison of the values of the descriptor when obtained from the intensity channel and from the (a) R,G,B and (b) H,S,V channels using the same cells and bins per histogram.

the set of these histograms as:

$$\widehat{h_{color_j}} = \frac{1}{3} \cdot \begin{bmatrix} \vec{h}_j^H \\ \vec{h}_j^S \\ \vec{h}_j^V \end{bmatrix} \tag{22}$$

Finally, the descriptor with the color features includes the set of normalized histograms of all the horizontal cells in which the image is divided. If n is the number of cells, we define the color histograms (CH):

$$CH = \frac{1}{n} \cdot \begin{bmatrix} \widehat{h_{color_1}} \\ \widehat{h_{color_2}} \\ \cdots \\ \widehat{h_{color_n}} \end{bmatrix} \tag{23}$$

In the same way, the descriptors of the spatial distribution are normalized. We denote them by $D_{spatial}$ independently on the method used to obtain them (FS, HOG or *gist*). In the case of the descriptors based on FS, the normalization is carried out by dividing each row by its first component in the frequency domain, which corresponds to the average value of the row. It should be noted that this value is different for each row of the transformed image. The normalization of the descriptors based on HOG and *gist* is carried out by dividing the elements of the descriptor by the sum of all their values. Finally, the weighing of the color and the spatial information can be weighted differently to compose the final descriptor:

$$D_{composed} = \begin{bmatrix} w_{spatial} \cdot D_{spatial} \\ w_{color} \cdot CH \end{bmatrix} \tag{24}$$

where $w_{spatial}$ and w_{color} are weighting factors.

This work includes a complete and systematic comparison of the different descriptors based on the global appearance described in Section II applied to panoramic images, focusing on the utility of the color information. With this purpose, several options will be tested and compared in subsequent sections. Each description technique will be applied separately (a) to the grayscale image, (b) to each RGB channel, (c) to each HSV channel, (d) both to each RGB and HSV channels to compose a unique descriptor and (e) the vector CH is calculated and appended to each of the different descriptors as explained in this section.

IV. SETS OF IMAGES

This section presents the sets of images used to carry out the experiments. These sets have been captured by ourselves in different areas and offices of the second floor of the Innova building of the Miguel Hernández University and are accessible from [49], where we can find more information about the dataset, including bird's eye views of the capture points both of the training and the test sets, the dimensions of every room and their distribution in a floor plan. Specifically, the datasets include images from a corridor (1), three offices with different configurations (2,3,4) a library (5) and a conference room (6). A catadioptric system is used to capture the datasets. It is composed of a color camera (Imaging Source model DFK-21BF04) and a hyperbolic mirror (Eizoh Wide70) which captures omnidirectional color scenes, with 640 × 480 pixel resolution.

Two datasets have been captured to test the performance of the descriptors: the training and the test ones. About the training dataset, the capture points compose a regular 40 cm × 40 cm grid on the floor, and all the captures are performed under real operating and lighting conditions. It is a challenging environment due to the presence of large windows that force us to reduce the gain of the camera to avoid the saturation of the image. For that reason, the histograms of the scenes are normally concentrated on the low area of the color range. Table 2 shows the number of images per area in the training dataset.

TABLE 2. Number of images included in each area of the training dataset.

Area	Number of Images
(1) Corridor	212
(2) Office 1	35
(3) Office 2	72
(4) Office 3	84
(5) Library	169
(6) Conference Room	300
Total	**872**

Second, the test dataset is composed of some images captured in the same environment, and they will be used to carry out experiments of position and orientation estimation. While capturing the test images, 3 different cases were considered about the capture points with respect to the training grid: (1) the test image is captured very close to the position of a training image; (2) the test image is captured halfway between two map images and (3) it is captured approximately equidistant to four images of the grid. In the experimental part (section V), the descriptors are evaluated in an image retrieval framework, in which the descriptor of each test image is compared with the descriptors of the training images and the most similar descriptor (nearest neighbour) is retained. In these experiments, the result will be considered a correct retrieval if the nearest neighbour was captured in the geometrically nearest point in case (1); in one of the two nearest points in case (2) and in one of the nearest 4 points in case (3).

These test images have been captured at different times of the day and days of the year, under real operating conditions, what hinders this task. This way, the test images include perceivable changes in lighting conditions, in the position of some pieces of furniture with respect to the training images and some people appearing in the scenes. These facts make the database more challenging.

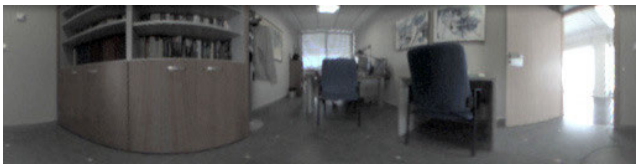
Additionally, from each test position, 16 different images were captured, with different orientations in the ground plane, with a lag of 22.5^o between consecutive rotations. Table 3 shows the number of test images per area.

TABLE 3. Number of images per area in the test dataset.

Area	Number of Images	Rotations	Total
(1) Corridor	12	x16	192
(2) Office 1	9	x16	144
(3) Office 2	10	x16	160
(4) Office 3	13	x16	208
(5) Library	16	x16	256
(6) Conference Room	17	x16	272
Total	**77**	**x16**	**1232**

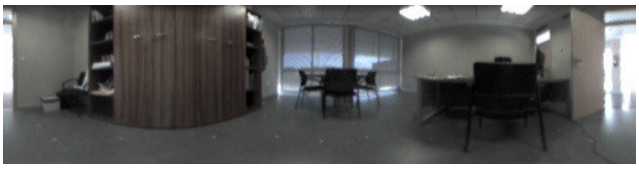
The descriptors included in this work are defined to be used with panoramic images. For that reason, we obtain the cylindrical projection of the omnidirectional images. Finally, the panoramic scenes are obtained by changing the cylindrical system to Cartesian coordinate system. The resolution of the panoramic images is 128×512 pixels. Fig. 8 includes a sample image from each area.

(a) Corridor



(b) Office 1

(c) Office 2



(d) Office 3

(e) Library

(f) Conference Room

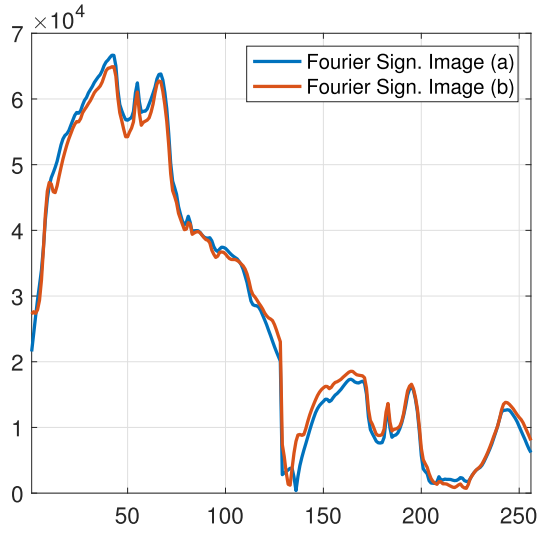
FIGURE 8. Examples of images captured from each area of the datasets.

It should be noted that, since it is an office environment, there are several elements that appear repeatedly in the different rooms with similar appearance. For that reason, the images might present visual aliasing. In that case, the descriptors may lose their capacity of distinguishing images due to the existence of similar scenes, and one of the objectives of the experiment is to check if any description method is able to cope robustly with this phenomenon.

As an example, Fig. 9 shows two scenes from the corridor, which are captured from two different positions, with a distance of 240 cm between them. We can see that their appearance is very similar. Fig. 9 (c) includes the Fourier Signature of both images. We can see that both descriptors are very similar despite the fact the scenes are different and separated.

(a) Corridor X=40, Y=280.

(b) Corridor X=40, Y=520.



(c)

FIGURE 9. Example of visual aliasing.

Additionally, the robustness of the descriptors is also tested when partial occlusions or noise appear on the test images to check if they could be able to operate if these phenomena occur in real-operation situations. The occlusions are introduced with four vertical stripes with different width that cover different percentages of the panoramic images. Regarding the noise, zero-mean Gaussian noise with different variances is artificially added to the different color channels.

Fig. 10 presents a panoramic image with examples of the occlusions, that vary from 5% to the 40% of the image, and with Gaussian noise, whose variance takes values between $\sigma = 0.0025$ and $\sigma = 0.0200$. They constitute specially challenging situations for the methods.

V. LOCALIZATION FRAMEWORK AND EVALUATION

The main objective of the paper is carrying out a comparative evaluation of the descriptors presented so far in a localization framework, focusing on the relevance of color information. This section is structured as follows. First, subsection V-A presents the localization framework implemented to carry out the tests and the measurements used to check the performance of the descriptors. Then, subsection V-B details the main parameters of the descriptors, whose sensibility is studied along the experiments.

A. ESTIMATING THE POSITION OF THE ROBOT

As outlined previously, the localization problem is addressed as an image retrieval problem. First, the descriptor of each training image is obtained. This set of descriptors is considered as the map. Second, for every test image, its descriptor is obtained and compared with all the descriptors stored in the map. The descriptor that presents the minimum Euclidean distance (nearest neighbour) is retained. This association is considered to be correct if the retrieved image is the one which was captured from the geometrically closest point. The performance of this process is evaluated throughout this work using two representations: recall-precision curves, and geometric distance between the capture point of the test image (ground truth) and the capture point of the retrieved map image.

Recall-precision curves [50], [51] permit evaluating the performance of the descriptors in image association tasks. The concepts of *recall* and *precision* are defined as:

$$recall = \frac{\#\text{ of correct matches retrieved}}{\#\text{ total of correct matches}} \quad (25)$$

$$precision = \frac{\#\text{ of correct matches retrieved}}{\#\text{ correct matches}} \quad (26)$$

This way, *recall* represents the ability of the descriptor to find all the correct associations, and *precision* the ability to find the correct associations as the number of experiments grows. Their values are between 0 (that would indicate that no correct match has been retrieved) and 1 (that would mean that the descriptor has found all the correct matches).

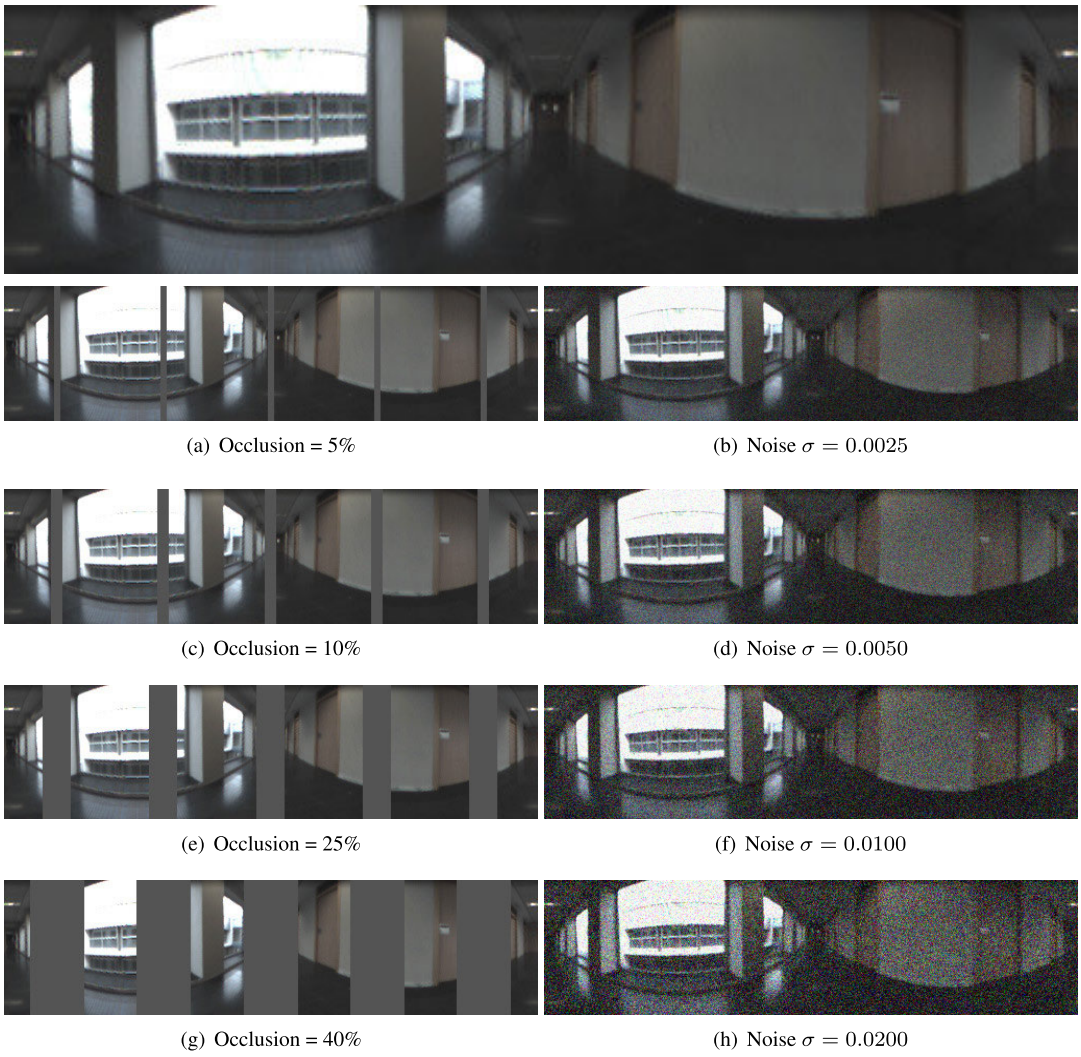
The process to obtain the recall-precision curves is the following:

1) First, we calculate the image distance between the test image descriptor and all the descriptors in the map. We define $\vec{d}_{Test} = \left[d_{Test,1}, d_{Test,2}, \ldots, d_{Test,n}\right]^T$ as the descriptor of the test image, which is a column vector with n elements, and $\vec{d}_i = \left[d_{i,1}, d_{i,2}, \ldots, d_{i,n}\right]^T$ the position descriptor of the i-th image of the map. The image distance is defined as:

$$l_{Test,i} = dist(\vec{d}_{Test}, \vec{d}_i) = \sqrt{\sum_{j=1}^{n}(d_{Test,j} - d_{i,j})^2}, \quad i = 1, \cdots, N \quad (27)$$

where N is the number of images in the map. After this step, an array of distances is available $\vec{l}_{Test} = \left[l_{Test,1}, l_{Test,2}, \ldots, l_{Test,N}\right]$.

2) The match between the test image and the map is determined from the minimum image distance in the vector $\vec{l}_{Test}$. Once the algorithm has calculated the Euclidean distance between the test image and all the images of the map, it selects the association with the minimum distance.



(a) Occlusion = 5% (b) Noise $\sigma = 0.0025$

(c) Occlusion = 10% (d) Noise $\sigma = 0.0050$

(e) Occlusion = 25% (f) Noise $\sigma = 0.0100$

(g) Occlusion = 40% (h) Noise $\sigma = 0.0200$

FIGURE 10. Test image including different levels of occlusion and Gaussian noise.

3) The algorithm determines whether the match is correct by checking if the capture point of the retrieved image is the one which is metrically closest to the capture point of the test image (ground truth).
4) After repeating this process for all the test images (N_{Test} is the number of test images) we obtain a matrix with N_{Test} rows and two columns. The first column contains the minimum Euclidean distance of each test image (min $\vec{l}_{Test}$), and the result of the match (1 or 0 depending on whether it is correct or not).
5) Then, we sort the association list in ascending order using the image distance, and obtain the values of *recall* and *precision* according to (25) and (26).

The distribution of the recall-precision curves provides information about the robustness of the descriptors with false positives considering a threshold in the image distance. So that, it is desirable that the *precision* keeps near 1 for every *recall* value, since it means that we have fewer false positives under that threshold. As an example, Fig. 11 shows two different recall-precision curves. Although the final values are similar, the distribution of the blue curve shows a better

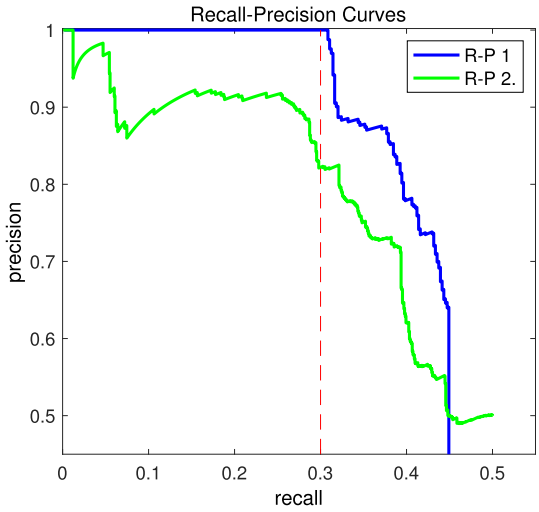


FIGURE 11. Comparison of two different recall-precision curves.

performance of the descriptor. If we set a threshold distance corresponding to *recall* $= 0.3$, the precision of *R-P 1* is 100%, but *R-P 2* is 82%. It means that we are able to obtain the

30% of correct matches with a 100% of probability in the first case, and with a probability of 82% in the second example.

To complement the results, we consider three different cases to create recall-precision curves, considering only the image with the minimum image distance (Nearest Neighbour or N.N.), looking for a correct match within the two cases with the lowest image distance (Second Nearest Neighbour or S.N.N.), or the three cases (T.N.N.).

We consider that it is interesting to analyze S.N.N. and T.N.N. apart from N.N. for the following reason. In this paper we focus on the role of color in the performance of the descriptors. Therefore, we address the localization task in a straightforward way: as a global localization problem, solved with an image retrieval approach (among all the images in the dataset). This assumes that we have no information about the previous pose of the robot while solving the problem. Only visual information is used. However, nowadays, many local localization algorithms exist (i.e. probabilistic algorithms) that take it into account the pose of the robot in the previous time instant to estimate the pose in the current time instant (apart from the odometry and visual information). In such algorithms, not only the N.N. but also other k-NN neighbors could play an important role, owing to the fact that the previous pose is known and the new pose should be at a relative distance form it.

B. PARAMETERS OF THE DESCRIPTORS

In Section II, we detail each of the compression techniques based on the global appearance that we include in this comparison. Next, we present a summary of the parameters of each one.

Regarding the descriptors based on the DFT, it is possible to select the number of elements of the transform in the frequency space. In the case of the 1D-DFT, the parameter is the number of elements retained from the transformed sequence. In the case of 2D-DFT, we can select the size of the submatrix that gathers the lower frequencies of the image. Finally, as for the FS, the parameter is the number of elements kept from every row. We select separately the number of elements retained from the magnitudes' matrix (N_{pos}), that allow us to carry out the estimation of the position of the robot in the map, and the number of elements in the arguments' matrix (N_{rot}), that provides information to estimate the orientation.

The technique that applies PCA over the FS is determined by the number of elements per row retained from the magnitudes' matrix (N_{pos}), and the number of main eigenvectors that compose the new projection basis (V_{PCA}). The orientation works as in the case of FS (N_{rot}). About rotational PCA, the main parameter of the descriptor is the number of rotations per image (R_{im}), and the number of eigenvectors that compose the new basis (V_{PCA}).

The parameters of HOG are the number of horizontal cells (cells with the same width that the panoramic image, denoted as C_H) used for localization purposes, and the width of vertical cells (cells with the same height of the image, denoted as S_V) as well as the distance between these vertical cells for the estimation of the orientation (D_V). Horizontal cells have no overlapping, so its number is determined by their height and the number of rows of the image. However, vertical cells may have some overlapping if the distance between consecutive cells is lower than their width. In the case of HOG, the number of bins per histogram is another parameter. However, we set this parameter to 8 since preliminary experiments showed that increasing the number of bins, the precision does not improve, but if it is lower, the precision decreases.

The parameters of *gist*-Gabor are the number of masks the image is filtered with, and the number of cells used to divide the filtered images. About the localization descriptor, we use two spatial scales for Gabor filtering in order to limit the computational cost. The variables are the number of masks considered in each of the two spatial scales ($Masks_1$ and $Masks_2$), and the number of horizontal cells that divide each filtered image (C_H). The filtering direction of the Gabor masks depends on the number of masks, since they are equally distributed between 0° and 180°. For the orientation descriptor, we only use the information in the first level of Gabor spatial filtering, with a maximum of 4 masks. So, for orientation, the parameters that define the descriptor are the width of the vertical cells (S_V) and the distance between them (D_V).

Finally *gist*-color uses always the same number of Gabor masks to filter the image, as stated in Section II, with 4 orientations. The filtering spatial scales are determined by the number of scales of the Gaussian pyramid. When a new image arrives, we create a pyramid with six levels. For the Gabor filtering, we use the three first levels of the pyramid. Regarding the color features, it uses the six levels to carry out the comparison between opposite color channels, as shown in Table 1. The parameters of the position descriptor are the number of horizontal blocks used to blockify the information of Gabor-filtered images, denoted as C_{HG}, and the number of cells used to blockify the information of color (C_{HC}). For orientation, we use only the information of the Gabor masks. The parameters are the number of vertical cells (S_V), and the distance between them (D_V).

In Table 4, a summary of the different parameters of each descriptor is included.

VI. RESULTS

The experimental section focuses on the comparative evaluation of the performance of the global-appearance descriptors with color information. We study the precision in the pose estimation (both position and orientation), using all the images in the test dataset, and comparing them with the map built from the descriptors of the training images (section IV). We also include a comparison of the computational time of each descriptor in the map building and pose estimation tasks.

This section is structured as follows. First, subsection VI-A carries out a study using the initial description methods presented in Section II (FS, HOG, *gist*-Gabor and *gist*-color) to adjust the different descriptor parameters, and to make a first comparison of computational requirements and performance

TABLE 4. Main parameters of the descriptors.

1D-DFT	Position	N_{pos}	Number of magnitude components
	Orientation	N_{rot}	Number of argument components
2D-DFT	Position	N_{pos}	Size of the magnitudes' submatrix ($N_{pos} \times N_{pos}$)
	Orientation	N_{rot}	Size of the arguments' submatrix ($N_{rot} \times N_{rot}$)
FS	Position	N_{pos}	Number of magnitude elements per row
	Orientation	N_{rot}	Number of argument elements per row
PCA over FS	Position	N_{pos}	Number of magnitude elements of the FS per row
		V_{PCA}	Number of eigenvectors selected after PCA analysis
	Orientation	N_{rot}	Number of argument elements of the FS per row
Rotational PCA	Position and Orientation	R_{im}	Equiangular artificial rotations of the image
		V_{PCA}	Number of eigenvectors selected after PCA analysis
HOG	Position	C_H	Number of horizontal Cells
	Orientation	S_V	Width of the vertical cells (pixels)
		D_V	Distance between consecutive vertical cells (pixels)
***Gist*-Gabor**	Position	$Masks_1$	Number of Gabor filter masks for the first scale
		$Masks_2$	Number of Gabor filter masks for the second scale
		C_H	Number of horizontal cells
	Orientation	S_V	Width of the vertical cells (pixels)
		D_V	Distance between consecutive vertical cells (pixels)
***Gist*-color**	Position	C_{HG}	Number of horizontal cells blocks of the Gabor images
		C_{HC}	Number of horizontal blocks of the opponent color channels
	Orientation	S_V	Width of the vertical cells (pixels)
		D_V	Distance between consecutive vertical cells (pixels)

of the descriptors. Then, subsection VI-B completes the experimental part including the use of color information as described in Section III. Finally, the performance of the descriptors when noise or occlusions are present in the test images is evaluated in subsection VI-C. All the algorithms and simulations have been developed using Matlab. The experiments have been performed using a computer with two Quad-core processors of 2.8GHz and 10GB of RAM. It is necessary to point out that it has not been possible running rotational PCA with the whole training dataset because of the excessively large computational requirements, specially RAM. For that reason, it appears with an asterisk in the graphs. The experiments of this descriptor use only three rooms of the training dataset, that correspond with the three offices, i.e. zones 2, 3 and 4 (Table 3). Only in the case of this descriptor, the reduced map is composed of 191 images, with 32 test locations, that means 512 test images considering their rotations. In the case of the other descriptors all the training and test images are considered.

In the remainder of this work we use the term Precision with the meaning stated in (26), and Accuracy to refer to the performance of the localization algorithms, as fas as geometric distance or orientation (measured in the ground plane between the pose of the test image and the pose of the nearest neighbour) are concerned.

A. RESULTS OBTAINED USING RAW DESCRIPTORS

In order to select the parameters of each descriptor and make a first study of feasibility, we consider only the initial descriptors (as presented in Section II) and the original space of representation of each technique. This space is the grayscale in the case of the descriptors based on DFT, HOG and *gist*-Gabor, and RGB in the case of *gist*-color. To tune the parameters of the descriptors, it is necessary to check both the performance in position and orientation estimation and the necessary calculation time. After a sensitivity analysis, the values selected for each parameter are shown in Table 5.

TABLE 5. Parameters selected for each descriptor.

Fourier 1D	Position	N_{pos}	32
	Orientation	N_{rot}	4
Fourier 2D	Position	N_{pos}	64
	Orientation	N_{rot}	8
Fourier Signature	Position	N_{pos}	32
	Orientation	N_{rot}	16
PCA over Fourier Signature	Position	N_{pos}	32
		V_{PCA}	872
	Orientation	N_{rot}	16
Rotational PCA	Position and Orientation	R_{im}	16
		V_{PCA}	100
HOG	Position	C_H	16
	Orientation	S_V	64
		D_V	4
***Gist*-Gabor**	Position	$Masks_1$	4
		$Masks_2$	8
		C_H	64
	Orientation	S_V	64
		D_V	32
***Gist*-color**	Position	C_{HG}	8
		C_{HC}	32
	Orientation	S_V	8
		D_V	16

The resulting *recall-precision* curves after solving the image retrieval problem with each of the different descriptors are shown in Fig. 12. They show that HOG and *gist*-color (Fig. 12(f) and 12(h)) show a better performance than the other descriptors. Moreover, the rates of false positives are very similar. The results of rotational PCA can also be highlighted, specially its low percentage of false positives until a relatively high *recall* value.

The final values of precision reached by FS (Fig. 12(b)), 2D-DFT (Fig. 12(c)) and *gist*-Gabor (Fig. 12(g)) are very similar, although *gist*-Gabor shows higher precision until a

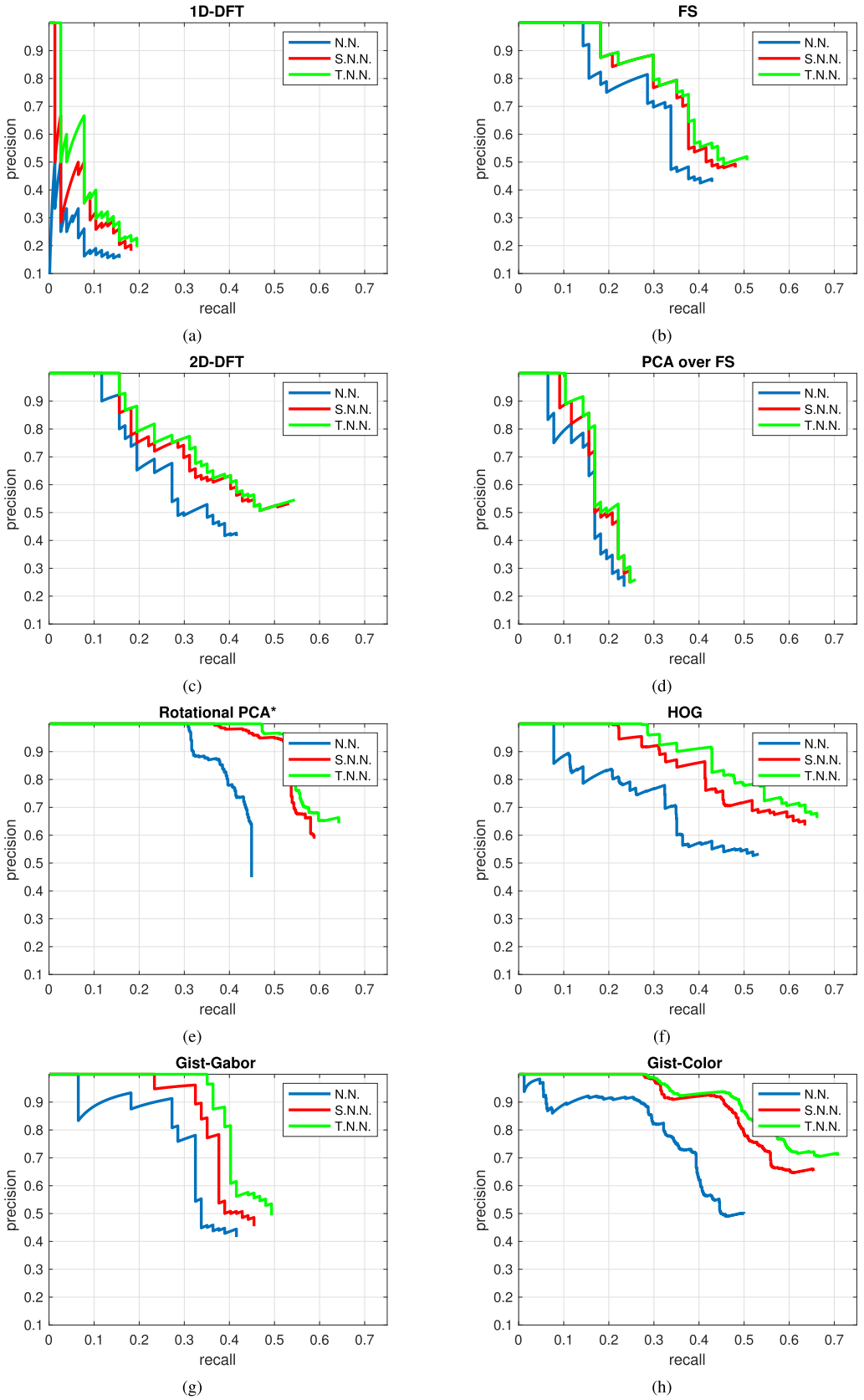


FIGURE 12. ***Recall-precision*** **results including three different measurements: the nearest neighbour (N.N.), the second nearest neighbour (S.N.N) and the third nearest neighbour (T.N.N.).**

recall value of 40%. 1D-DFT and PCA over FS (Fig. 12(a) and Fig. 12(d)) present the lowest rate of correct matches.

Next, Fig. 13 presents the accuracy of the position estimation process. The legend shows the average geometric distance between the capture point of each test image (ground truth) and the capture point of the nearest neighbour of the map (measured as the Euclidean distance on the floor). Therefore, the figure shows the percentage of experiments under a specific geometric distance. We can appreciate that the results have a similar behaviour as shown in the recall-precision curves, with HOG and *gist*-color the descriptors that present a better performance, specially HOG.

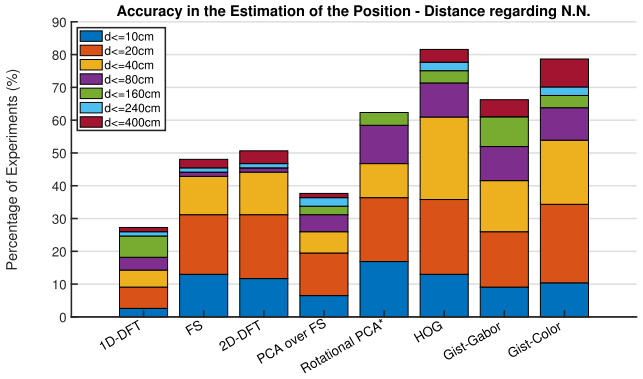


FIGURE 13. Accuracy in the estimation of the position. The legend shows the average geometric distance between the capture point of each test image (ground truth) and the capture point of the nearest neighbour of the map.

Although FS, 2D-DFT and *gist*-Gabor present final values of precision which are similar, the descriptor based on *gist* shows a better performance regarding the geometric distance to the image retrieved from the map, obtaining similar results than rotational PCA. It is important to highlight that, for some experiments, it is not possible to have an error $d \leq 10cm$, due to the resolution of the grid and the position of the capture points of the test images (e.g. some test images are in the middle of the 40 × 40 cm grid of the map). To know how significant these rates are, the size of the environment is shown in [49]. In order to understand the relative performance of each localization algorithm, the size of the whole environment must be considered, since the experiments include all the images of the map and they are not limited to individual rooms (i.e. the localization is approached as a global localization problem).

Regarding the orientation estimation, Fig. 14 presents the error in the estimation. To evaluate the performance of the descriptors in the orientation estimation, we consider only the associations whose geometric localization error is lower than 40 cm. That way, we avoid to estimate the phase lag between images that are too far from each other and the performance of each descriptor in orientation estimation is more realistically addressed, making this study more independent on the accuracy of the position estimation.

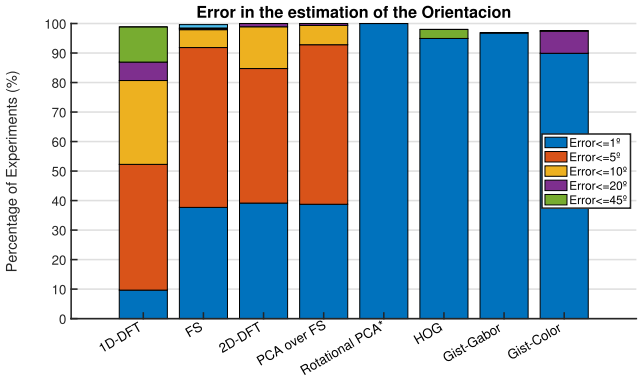


FIGURE 14. Error in the estimation of orientation. Only experiments whose geometric error in position estimation is ≤ 40cm are considered.

The techniques that present the best results in the orientation estimation are rotational PCA, *gist*-Gabor, *gist*-color and HOG. However, we should remark that in those descriptors, the phase information is sampled, either by the number of rotations of the images included in the map (in rotational PCA) or by the number of vertical cells.

The error in orientation estimation with FS and 2D-DFT is similar. FS and PCA over FS estimate the orientation using the same algorithm since PCA is not applied to the phase information. The slight difference between both results is consequence of the different associations during the position estimation. 1D-DFT presents the highest error in the phase estimation. Even so, it provides 80% of the experiments with an error equal or less than 10° using only 4 terms per image.

In these experiments, the map is composed of the descriptors of all the images from the different rooms. Fig. 15 shows the size of the map using the different descriptors, including the memory to store position and orientation information separately. The most compact descriptor is 1D-DFT, followed by HOG and *gist* descriptors. To improve the orientation accuracy in HOG and *gist*, the growth of the orientation descriptor size would be noticeable. Regarding rotational PCA, the memory requirements include the projection basis with the selected eigenvectors, the projection of the original map into the new basis, and the difference of phases between consecutive projections. As shown in Fig. 15, the information of orientation is insignificant compared to the location. However, to improve the accuracy in orientation estimation, more rotated siblings of each initial image should be considered and the computational cost of the mapping process would be even greater. Finally, we can see that after the projection to the new basis, the database of position estimation of FS is reduced from 29Mbytes to 4Mbytes when PCA is applied.

Fig. 16 shows the time spent in the map building and in the pose estimation of the robot, including both position and orientation. Regarding the map building (Fig. 16(a)) the techniques based on DFT can be considered the most efficient, except for PCA over FS, since PCA is a computationally expensive process, being 15 times greater than FS. Rotational PCA is the algorithm that spent more time in the

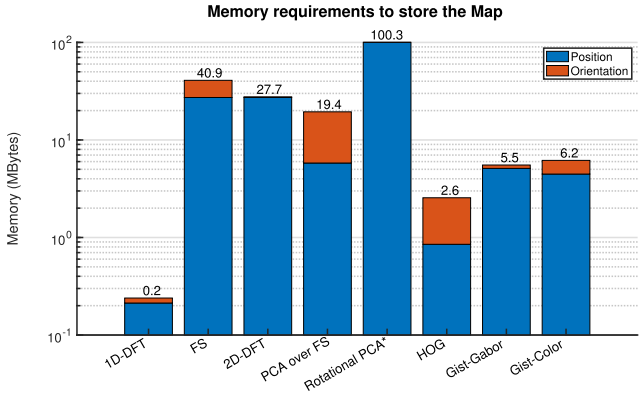


FIGURE 15. Memory requirements to store the map, showing separately the size of the position and orientation information.

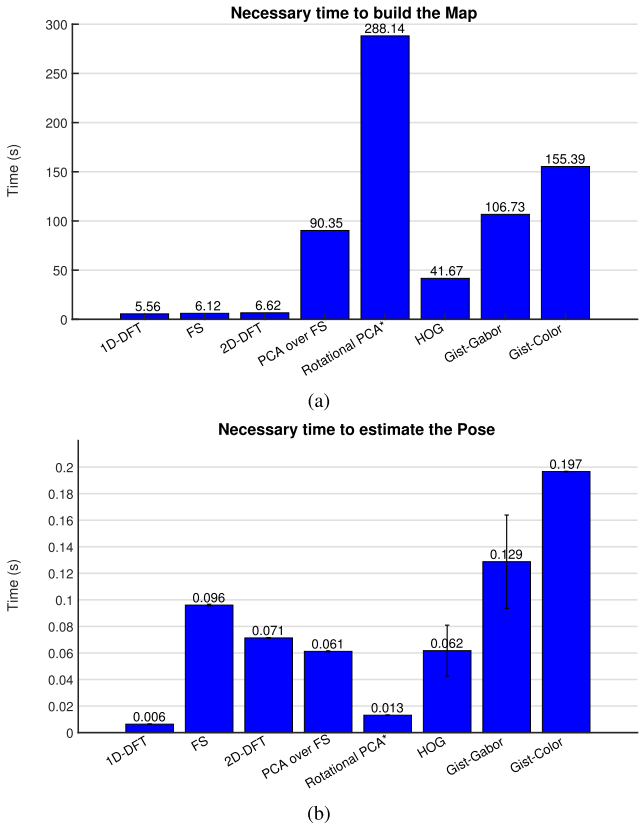


FIGURE 16. Time for (a) map building and (b) pose estimation.

map building task. Additionally, HOG and *gist* descriptors require more time than the descriptors based on DFT, specially *gist*-color.

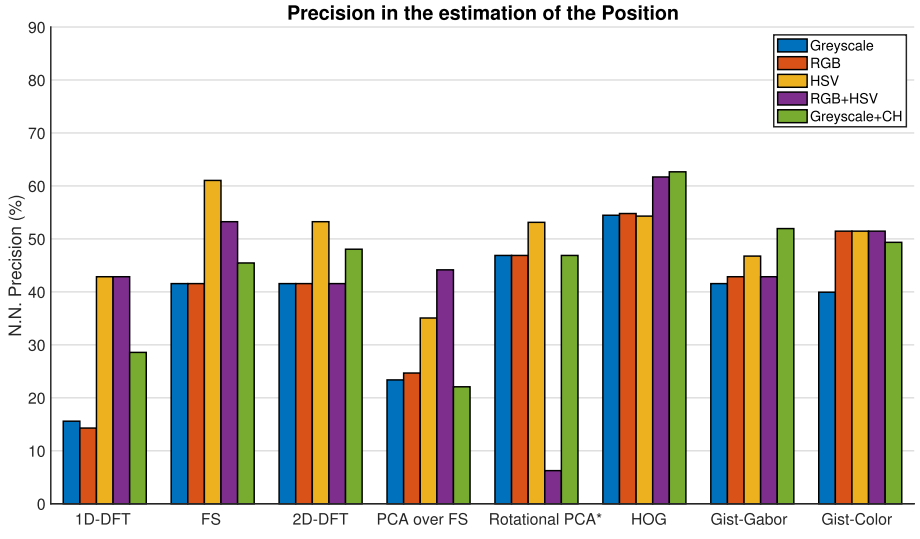
The pose estimation time, showed in Fig. 16(b), includes the necessary time to create the descriptor of the test image, the estimation of the position, and the orientation. HOG and *gist* descriptors present similar time values to create the descriptor. Compared to the other techniques, FS and 2D-DFT present a relatively high time in the estimation of the pose compared to the map creation. This is due to the estimation of the orientation, since it is a computationally complex process, specially in the case of FS. However, since 1D-DFT only uses 4 phase components, it is not affected by this fact. Finally, rotational PCA is one of the fastest algorithms in the pose estimation, since it only projects the image into the new basis, and calculates the image distance.

B. RESULTS OBTAINED USING COLOR INFORMATION IN THE DESCRIPTORS

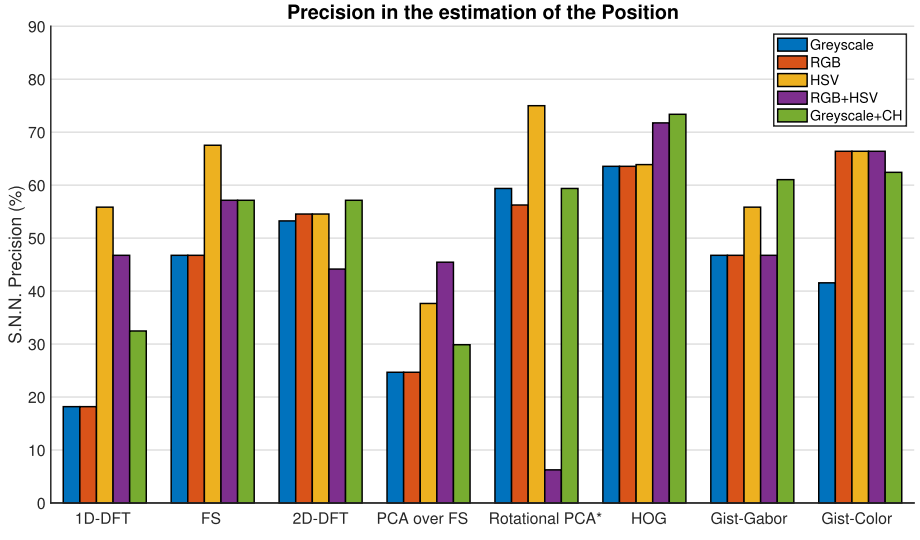
In this section, the results of the position estimation and computational requirements of each algorithm using the color information are presented. The comparison includes the application of the methods 1D-DFT, FS, 2D-DFT, PCA over FS, rotational PCA, *gist*-Gabor and *gist*-color to different color channels, to obtain a variety of descriptors. The performance of each descriptor will be tested subsequently in a position and orientation estimation task.

The next combinations are tested: (a) applying the description method to each RGB channel, to obtain a unique descriptor per scene; (b) the same to the HSV channels; (c) appending the descriptors obtained with (a) and (b) to create a unique descriptor; (d) applying the description method to the intensity channel and appending the color information using the Color Histograms (CH), as seen in (24). It is worth highlighting that the CH information has been used differently for each PCA based method, to adapt it to each description method. In the case of PCA over FS, first, the FS descriptor and CH are created, and then appended to form a vector. After that, PCA is performed with these data vectors. However, in the case of rotational PCA, the projection of the images in the new basis is obtained first, and then the vector CH is appended to this projection. In order to complete the experimental evaluation, we create a version of *gist*-color for grayscale space. The comparison among the color channels is replaced by the multiscale comparison in grayscale space. As far as the experiments are concerned, we use the parameters included in Table 5. For the Color Histograms, we use 8 or 16 horizontal cells depending on the descriptor, and 32 bins per histogram. In order to limit the number of variables when comparing the performance of the different descriptors and color spaces, in the combination of the greyscale descriptor and the Color Histogram (combination (d), denoted as Greyscale+CH in the experiments) we set the coefficients of the and color information, giving them the same weight ($w_{spatial} = w_{color} = 0.5$). At the end of the work, we include a study of the effect of varying these weighting coefficients in the different descriptors.

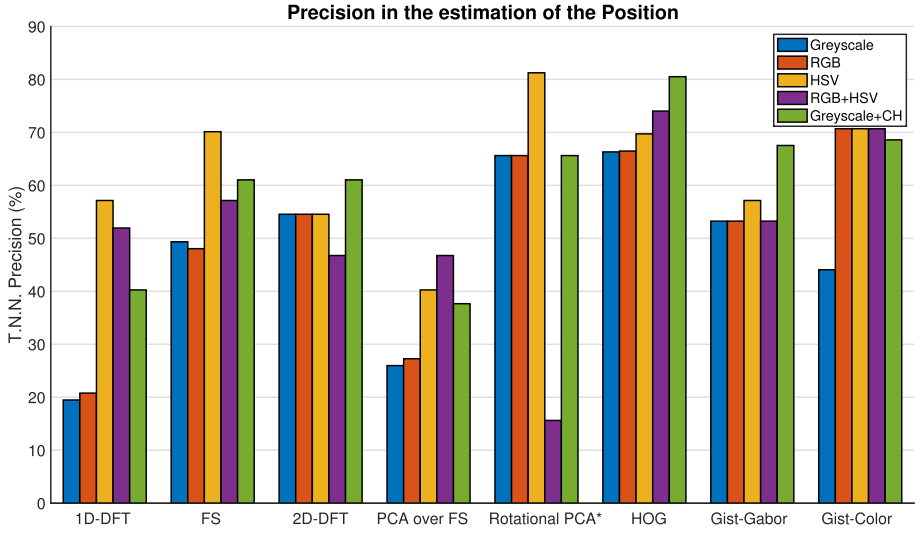
Fig. 17 shows the precision in the estimation of the position using all the combinations while building the descriptor. According to the results, the descriptors improve their performance substantially when the color information is used, except FS with RGB, 2D-DFT with RGB+HSV and rotational PCA using RGB+HSV. It is specially remarkable the improvement of 1D-DFT applied to HSV, since it triples its precision considering the T.N.N., from 19% to 57%. The performance of HOG with CH can be also highlighted, with a T.N.N precision over 80%.



(a)



(b)



(c)

FIGURE 17. Precision in the estimation of the position using the color information. Results considering (a) N.N. (b) S.N.N. and (c) T.N.N.

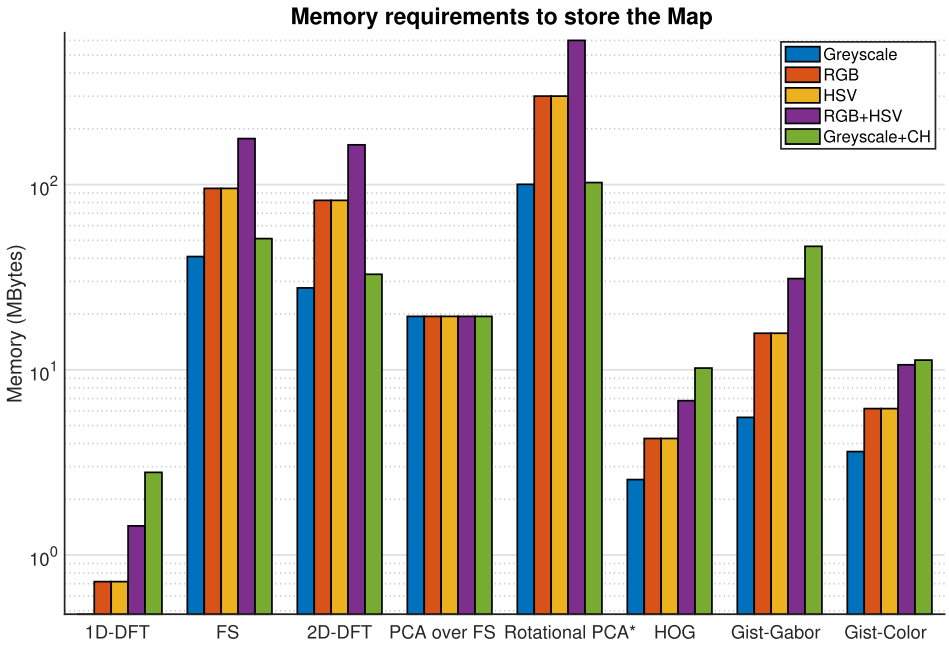


FIGURE 18. Size of the map when the color information is included in the descriptors.

The results of RGB color space show no significant improvements with any descriptor compared to grayscale. This is due to the high correlation between channels R, G and B, as stated in Section III. The exception is *gist*-color. As stated before, this descriptor is specially designed for the RGB space, since it includes the color opponency comparison of Hering (section II-D.2). When we add the information of Color Histograms, the performance of all the descriptors improves. 2D-DFT, HOG and *gist*-Gabor specially benefit from the addition of this color information.

On a general basis, the application of the descriptors over HSV channels presents the same or better results than over RGB channels. The improvement is specially significant in the case of 1D-DFT, FS and rotational PCA.

The necessary memory to store the map (position descriptors), using each combination, is presented in the bar graph included in Fig. 18. As expected, the use of RGB or HSV color spaces triples the memory of the map, and the joint use of both color spaces (RGB+HSV) multiplies by six. Regarding the Color Histogram, it adds a fixed quantity of information. PCA over the FS is the only descriptor that does not increase the size of the map using different color spaces, since PCA is applied to all the information that composes the original basis, and a fixed number of eigenvectors is selected in all the cases.

Fig. 19 shows the necessary time to build the map (Fig. 19(a)) and to estimate the pose of the robot in this map (Fig. 19(b)). As far as the map building task is concerned, PCA over FS and rotational PCA are the algorithms that take more time. This fact demonstrates again that PCA is a computationally expensive process, specially when the size of the map increases when using RGB and HSV color spaces. Last, *gist* techniques require, in general, more time than methods based on DFT or HOG.

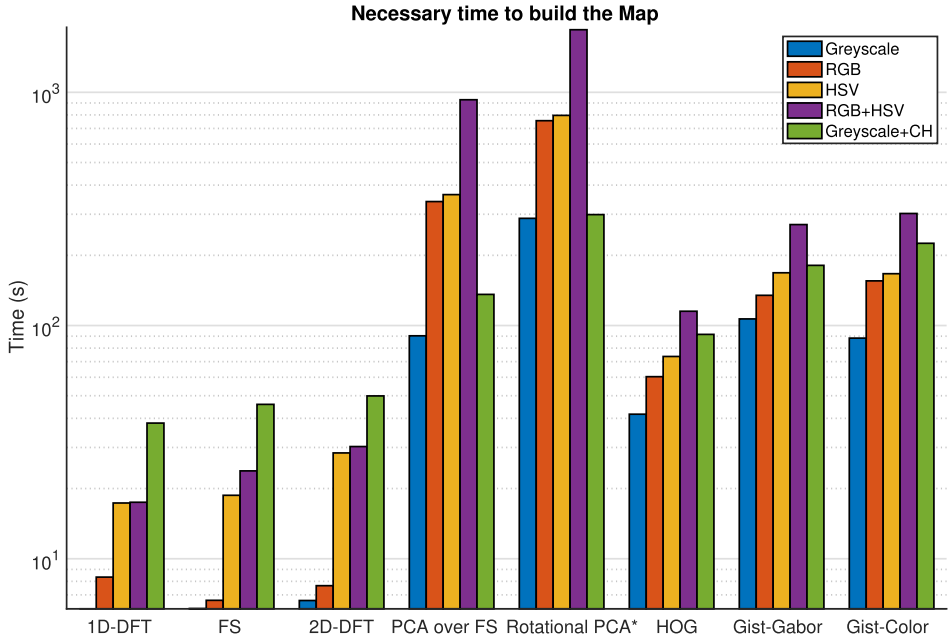
In the pose estimation task, methods based on PCA are remarkably fast. Additionally, except for 1D-DFT, all the techniques based on DFT use approximately the same time than *gist* methods. HOG is a descriptor with relatively low computational time. Regarding the color spaces, when we increase the number of color channels, the time rises, as expected. When we use descriptors over HSV, although the number of channels is the same than RGB, the time increases slightly due to the color space transformation of the original image.

The calculation of the Color Histogram varies between about 0.05 and 0.1 seconds depending on whether we use 8 or 16 cells per image. In the case of 1D-DFT, PCA over FS and rotational PCA, when we add the CH, the estimation of the pose takes more time than when we use the other color channels. For FS and 2D-DFT, the pose estimation time using the CH is only lower than RGB+HSV.

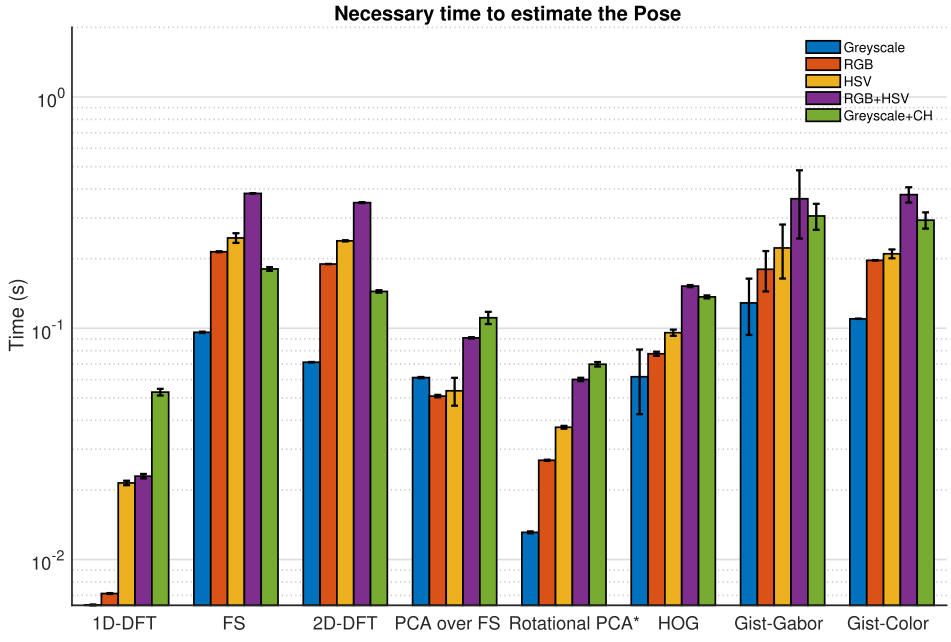
In general, the precision when we use the color information is higher than using only greyscale space. HSV and grayscale+CH are the methods that present the best results, except for PCA over FS, that achieves the best performance using RGB+HSV.

C. ROBUSTNESS AGAINST THE PRESENCE OF OCCLUSIONS OR NOISE IN THE TEST IMAGES

In order to complete this comparative evaluation, a set of additional experiments is performed to test the robustness of the descriptors. In these experiments, the effect of Gaussian



(a)



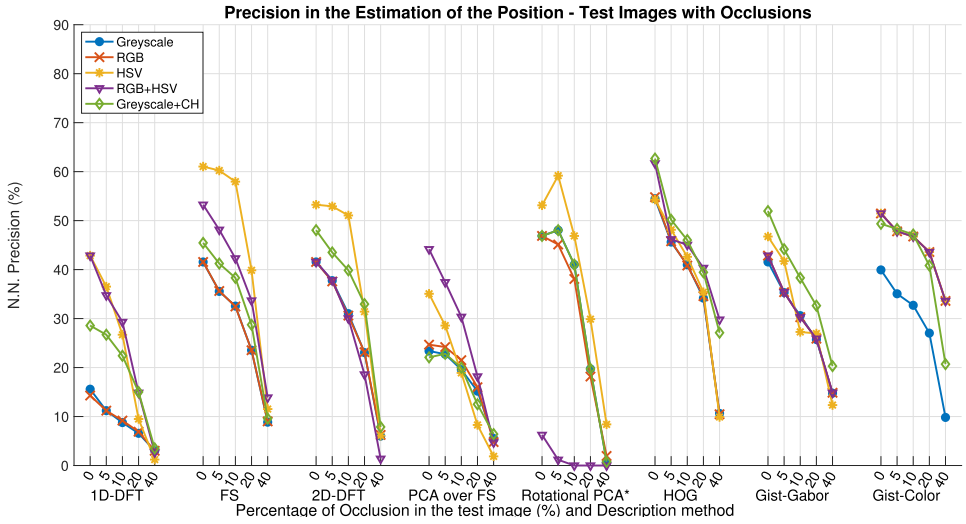
(b)

FIGURE 19. Necessary calculation time using the color information for (a) map building and (b) position estimation.

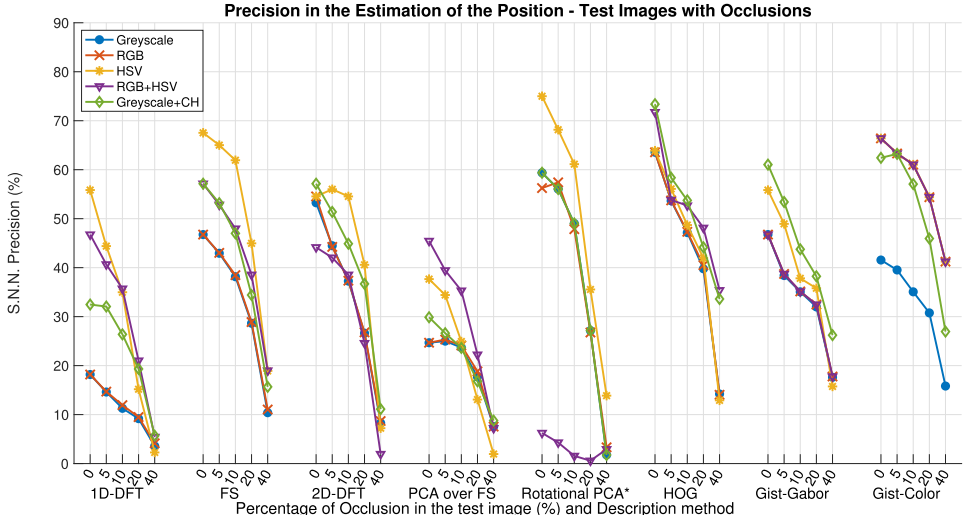
noise and partial occlusions in the test images is tested. These experiments allow us to check the performance of the different techniques under challenging, complex and changing environments, as it happens in real environments under real-operation conditions.

Fig. 10 shows a test image with different percentages of occlusion and with Gaussian noise with different variances. These challenging test images are used in this section to test the performance of the descriptors. Fig. 20 shows the results for the occlusion experiments. The horizontal axis shows the different description combinations considered in the analysis and, for each combination, the percentage of occlusion in each test image.

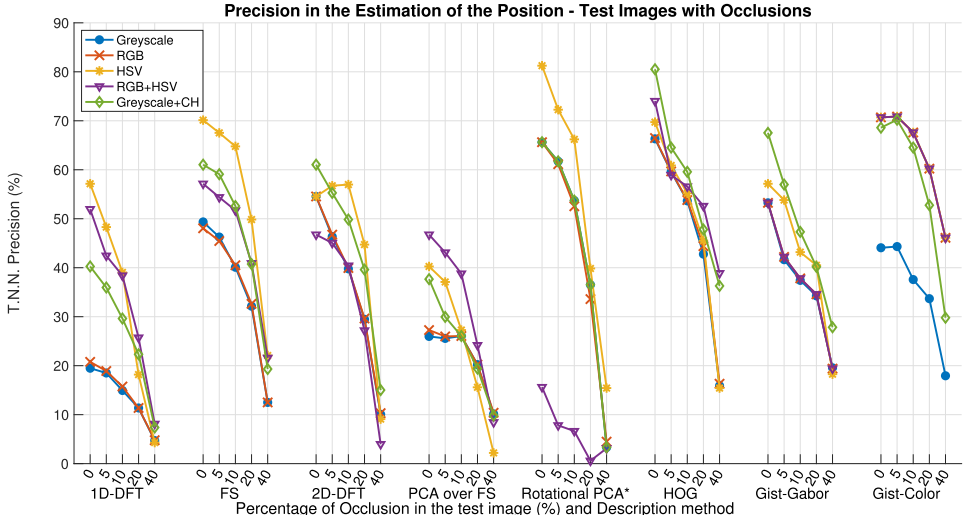
According to the results, the descriptors that are more negatively affected by occlusions are 1D-DFT and rotational PCA, specially the last one. HOG and *gist* descriptors are less



(a)

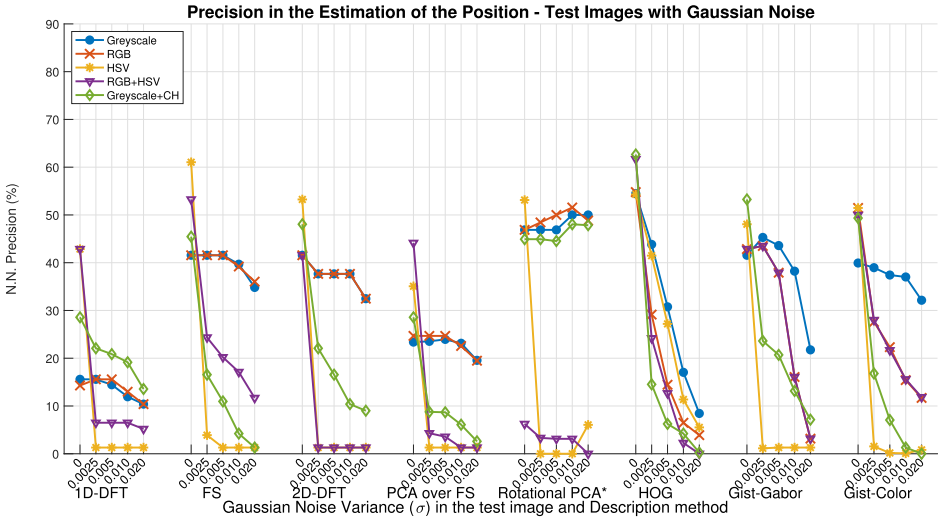


(b)

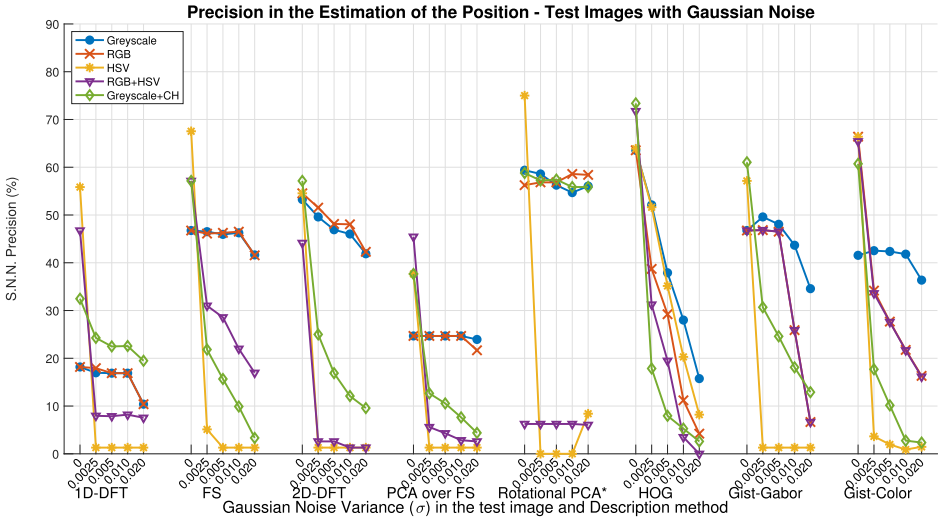


(c)

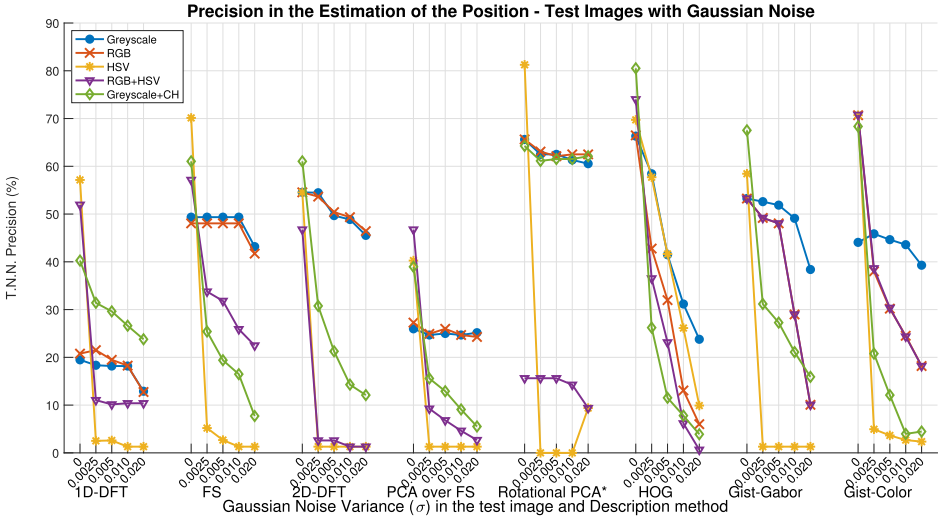
FIGURE 20. **Precision in the estimation of the position using the color information when the test images present occlusions. Results considering (a) N.N. (b) S.N.N. and (c) T.N.N.**



(a)



(b)



(c)

FIGURE 21. Precision in the estimation of the position using the color information when the test images present Gaussian noise. Results considering (a) N.N. (b) S.N.N. and (c) T.N.N.

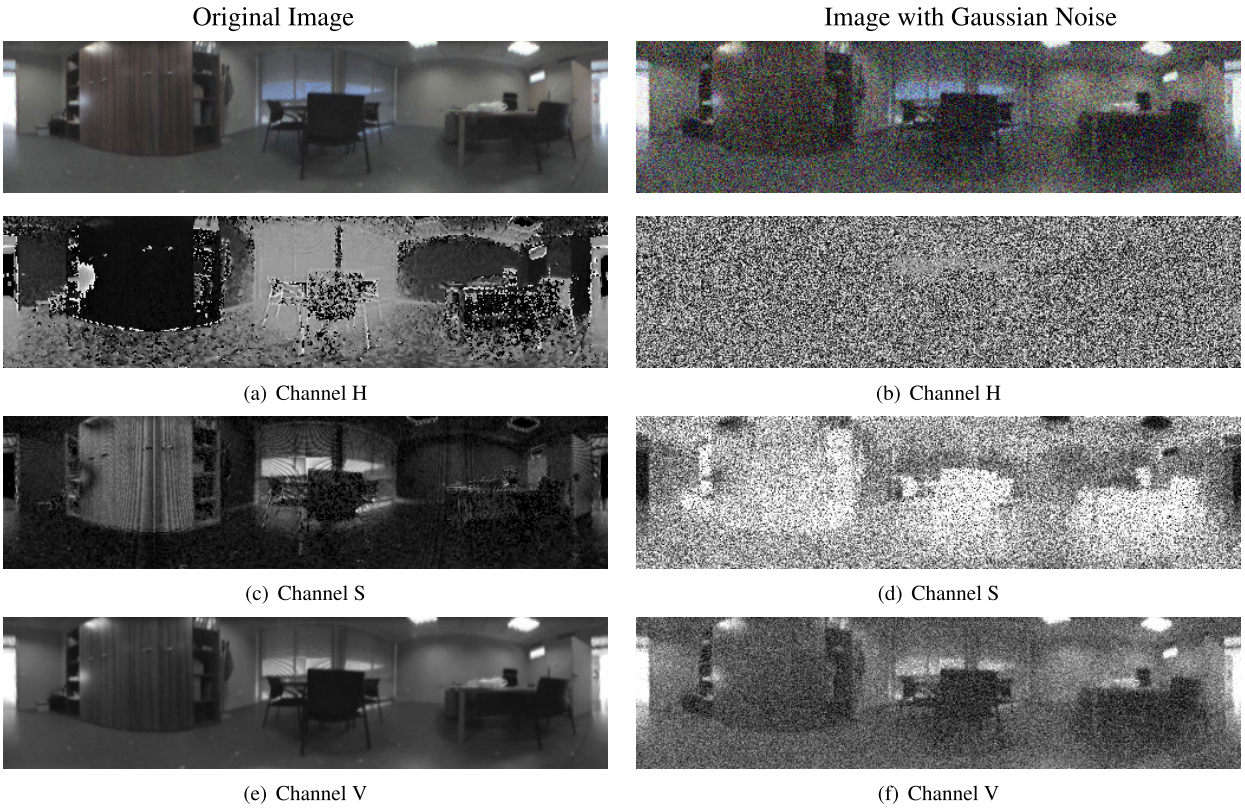
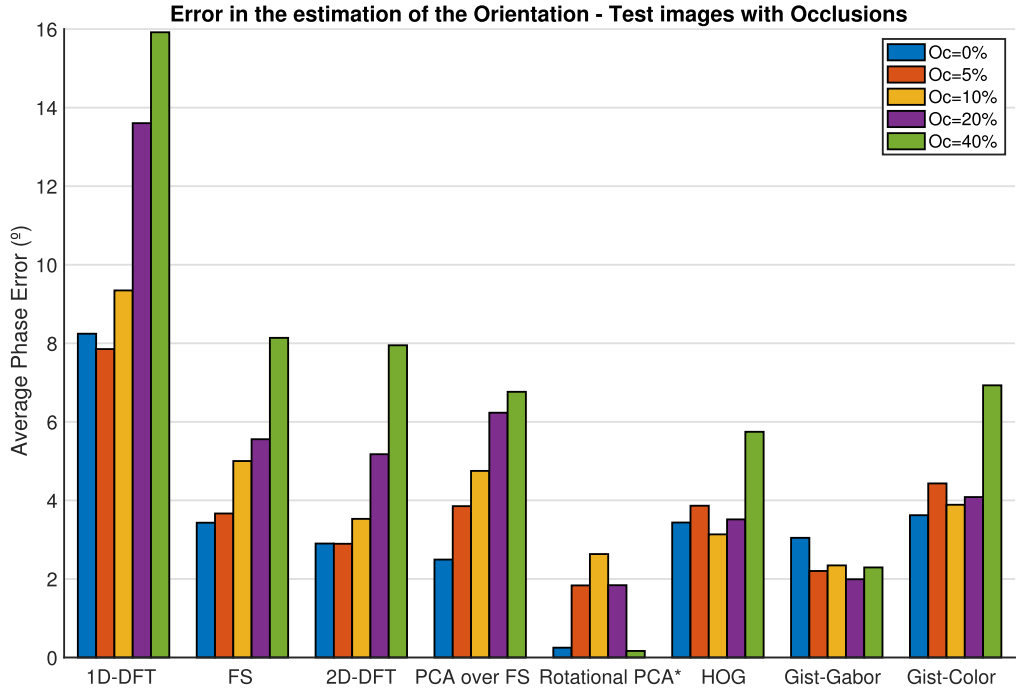


FIGURE 22. Channels of a sample image from the map in HSV space: original channels and with Gaussian noise.

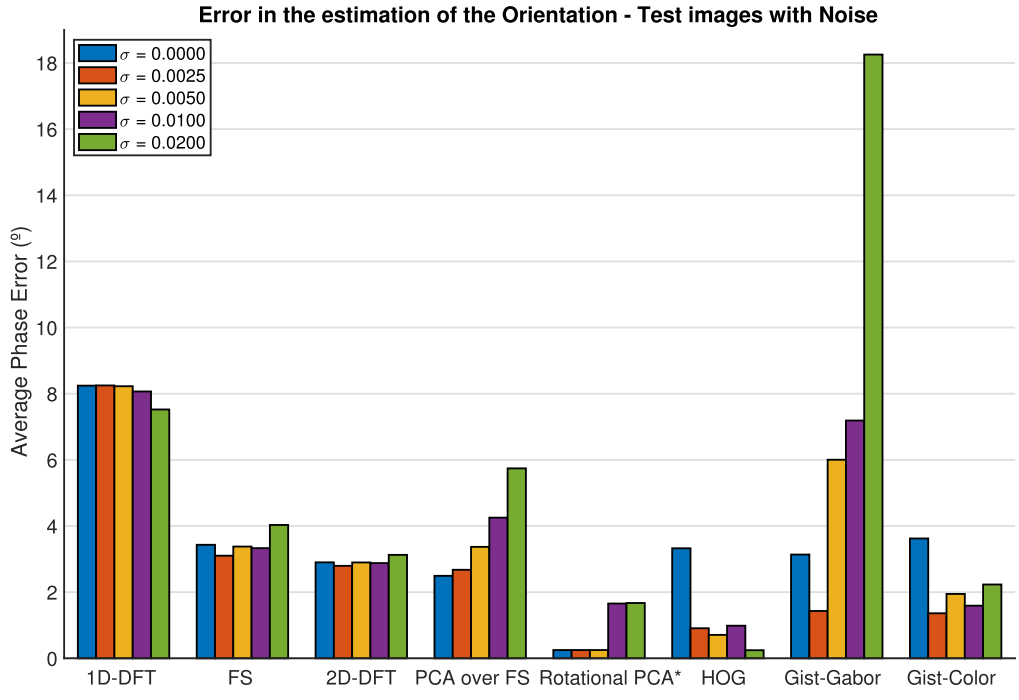
sensitive to occlusions in the image, specially HOG using Greyscale+CH, and *gist*-color using any color method. We can highlight the performance of FS, 2D-DFT and *gist*-color up to 10% of occlusion in the test image. On a general basis, the higher percentage of occlusion, the lower precision, as expected. However, some descriptors present a relatively good behaviour for occlusion percentages up to 10%, such as FS with HSV, 2D-DFT with HSV and gist-color.

Additionally, Fig. 21 shows the performance of the descriptors when the test images are affected by different levels of Gaussian noise. The horizontal axis shows the different description combinations considered in the analysis and, for each combination, the variance of the Gaussian noise which is present in each test image. We can see that the HSV color space is specially sensitive to Gaussian noise. Only HOG along with HSV presents a precision which is similar to the precision obtained with other color spaces. To obtain the images with noise, the Gaussian noise is added to R, G and B channels of each original test image separately. Fig. 22 shows the channels H, S and V of a test image without noise and with Gaussian noise with mean equal to zero and $\sigma = 0.0200$. We can observe that channels H and S are especially affected by noise, doing it almost impossible to recognize the original image. When the descriptors use the Color Histograms, the results present a reduction of the performance when the image present noise. Comparing the results of grayscale and grayscale+CH, the CH improves the location precision only with 1D-DFT. As far as the description techniques are concerned, FS, 2D-DFT and rotational PCA present a better performance.

Next, we include the results in the estimation of orientation when the test images are affected by occlusions (Fig. 23(a)) and by Gaussian noise (Fig. 23(b)). As in subsection VI-A, the error in the estimation of the orientation is only calculated in the experiments whose position error is equal or less than 40 cm in the map. Moreover, the orientation information is calculated only from the greyscale space. First, regarding occlusions, methods based on the DFT present a significant increase of the error. Results show that, for 40% occlusion, the error doubles compared to the original test images (with no occlusion). This is specially significant in the case of 1D-DFT. Rotational PCA shows the best performance in the orientation estimation. In the case of *gist*-Gabor, the average error is below 3° for any occlusion level of the test image. HOG and *gist*-color present a similar precision, with a mean error lower than 8°. Second, about the presence of Gaussian noise, the most affected descriptors are *gist*-Gabor and PCA over FS. The orientation error in *gist*-Gabor is specially high, multiplying by 6 the mean error when the noise variance is 0.020. Techniques based on DFT (except for PCA over FS) present little variation when the test image is affected by noise, with a similar error in all the cases. Finally, HOG and *gist*-color present a better performance in the orientation estimation when the image presents noise than with occlusions.



(a)



(b)

FIGURE 23. Orientation error when the test images present (a) occlusions and (b) Gaussian noise.

Finally, an experiment has been conducted to study the effect of considering different weighting factors in (24) when combining the greyscale descriptor and the Color Histogram (denoted as Greyscale+CH). Figure 24 includes the precision in localization for all the descriptors using Greyscale + CH when the weighting coefficients vary. On a general basis, the precision is higher when the combination $w_{spatial} - w_{color}$ is $0.5 - 0.5$ or $0.6 - 0.4$. However, in 1D-DFT we can see that the results improve when w_{color} increases. The reason is that this spatial descriptor is very compact and contains little

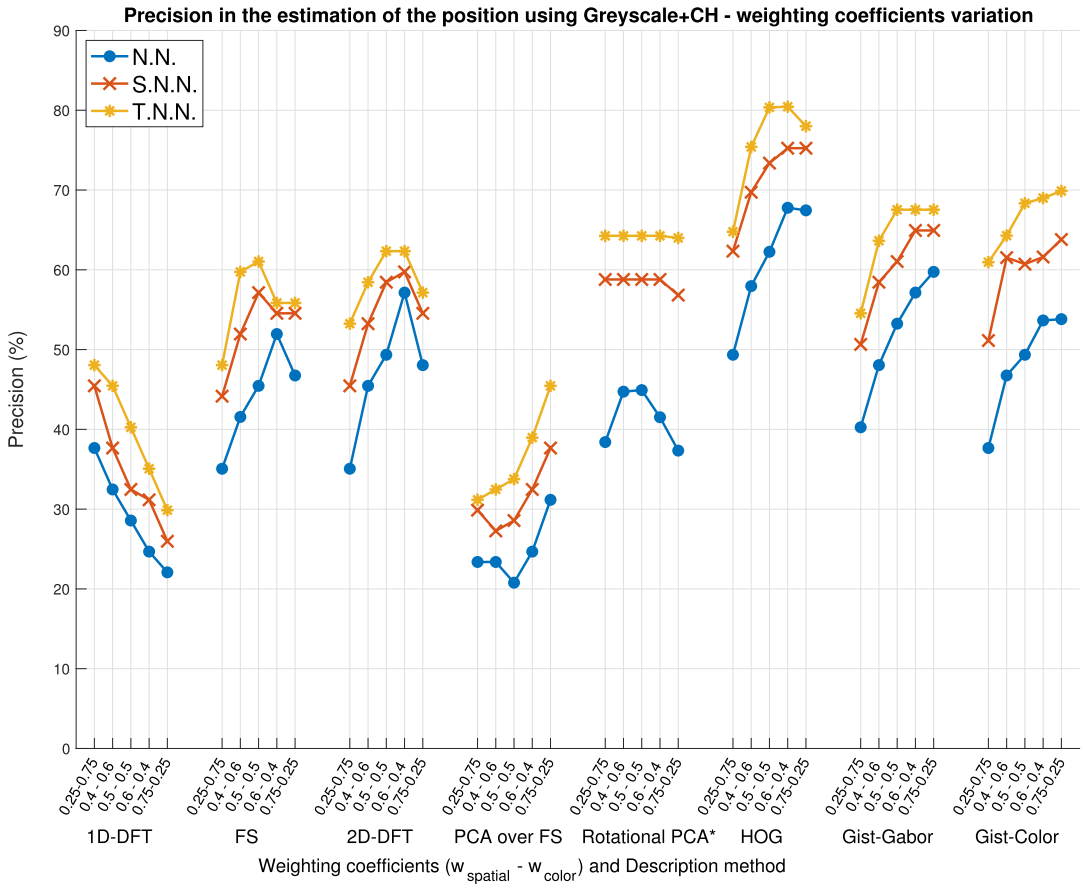


FIGURE 24. Precision in the estimation of the position for Greyscale + CH space varying the weighting coefficients.

information about the scene. On the contrary, *gist* descriptors and PCA over FS present a better performance when $w_{spatial}$ is higher, specially in the Nearest Neighbour experiments.

VII. CONCLUSION

In this work, the role of color information in the construction of global-appearance descriptors has been explored. A complete comparative evaluation has been carried out to uncover the performance of a variety of description methods and color channels in a pose estimation task. This evaluation has included the calculation time and the memory requirements of the different descriptors in the creation of a dense map, and the time consumed and precision in the estimation of the position and orientation of a robot in this map. Also, the robustness of these methods against the presence of noise or occlusions has been studied.

Next, we gather the main conclusions of this evaluation:

Position estimation and computational requirements

- In general, the use of the color information improves the performance of the descriptor in the localization task. This happens with all the description techniques.
- The color space HSV provides better results than RGB, except when the test images are affected by Gaussian noise.
- When the Color Histogram information is appended to the descriptors, the percentage of correct localization improves (comparing it with the case of using only the information in the grayscale space).
- Except for the FS and 2D-DFT, the computational cost that supposes appending the Color Histogram information is higher than if the color information is obtained by applying the descriptor over RGB and/or HSV channels.
- The combined use of RGB+HSV does not mean an improvement in the localization performance compared to the use of only a color space (RGB or HSV), but it supposes a significant increase of the computational requirements (both time and memory).
- Rotational PCA presents high precision in the pose estimation task, although the computational requirements make it infeasible for the task of dense map building in large environments. Moreover, together with PCA over FS, they are the only techniques that do not permit us building the map incrementally (i.e. all the images to be included in the map must be available initially). When a new image must be added to the map, PCA must be carried out with all the images (including the new one) from the scratch.

- HOG presents a good trade-off between precision and computational requirements, specially when we introduce the Color Histogram to the descriptor. Moreover, it is a very compact descriptor.
- 1D-DFT is the most compact descriptor using any color space. Although the precision of position estimation is low when we use only the grayscale image and the RGB color space, the precision increases to 58% when we use HSV. It becomes an interesting descriptor if the algorithm application has important restrictions of time, and specially, memory.
- The FS and 2D-DFT present a reduced computational cost during the map building process. However, the necessary memory to store the map is high compared to the rest of descriptors, and so the calculation time for the pose estimation is, because of the orientation estimation algorithm.
- *Gist*-color shows a better performance in the localization task than *gist*-Gabor, although it also needs more time during the map building and pose estimation processes.
- The *gist* descriptors need more time than the descriptors based on the DFT in the map building process, although they lead to maps whose size is substantially lower.

Estimation of the orientation

- Regarding the estimation of the orientation, all the descriptors present an average error lower than 8° when the test image is not affected by occlusions or noise.
- The DFT-based techniques produce a higher error in orientation estimation compared to the other description techniques.
- However, it is worth noting that the angular resolution depends directly on the information included in the descriptor in the case of *gist*, HOG and rotational PCA, since it is sampled. This resolution can be increased at the expense of increasing the size of the descriptor and the calculation time both to build the map and to solve the orientation estimation problem. Therefore, in those descriptors, the phase information is less flexible and more sensitive than the descriptors based on DFT.

Localization when the test images present partial occlusions

- In general, the effect of the occlusions is more significant when we use the color spaces than using grayscale images. However, the different descriptors still present a better performance using color spaces than greyscale space.
- *Gist* and HOG are the techniques which are less affected by occlusions. The results obtained with the combinations HOG using grayscale+CH, and *gist*-color over RGB and HSV are specially remarkable.
- In the orientation estimation, the DFT-based descriptors are the least robust against the presence of occlusions in the test images, specially 1D-DFT. However, except this descriptor, the average error remains below 8°.

Localization when the test images present noise

- The Gaussian noise remarkably affects the channels *Hue* and *Saturation* in the space HSV. This implies a significant reduction in the localization precision when we use the color space HSV and RGB+HSV.
- Descriptors based on the DFT, rotational PCA and *gist*-color present the lower reduction in the precision when noise is present in the test images.
- In the estimation of the orientation, only PCA over FS and *gist*-Gabor show an important increase in the error.

REFERENCES

[1] E. Garcia-Fidalgo and A. Ortiz, "Vision-based topological mapping and localization methods: A survey," *Robot. Auto. Syst.*, vol. 64, pp. 1–20, Feb. 2015.
[2] I. Ohya, A. Kosaka, and A. Kak, "Vision-based navigation by a mobile robot with obstacle avoidance using single-camera vision and ultrasonic sensing," *IEEE Trans. Robot. Autom.*, vol. 14, no. 6, pp. 969–978, Jun. 1998.
[3] Z. Zhu, S. Bhattacharya, M. Uijt de Haag, and W. Pelgrum, "Using single-camera geometry to perform gyro-free navigation and attitude determination," in *Proc. IEEE/ION Position, Location Navigat. Symp. (PLANS)*, May 2010, pp. 858–867.
[4] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 993–1008, Aug. 2003.
[5] Z. Yong-guo, C. Wei, and L. Guang-liang, "The navigation of mobile robot based on stereo vision," in *Proc. 5th Int. Conf. Intell. Comput. Technol. Autom. (ICICTA)*, Jan. 2012, pp. 670–673.
[6] Y. Jia, M. Li, L. An, and X. Zhang, "Autonomous navigation of a miniature mobile robot using real-time trinocular stereo machine," in *Proc. IEEE Int. Conf. Robot., Intell. Syst. Signal Process.*, vol. 1, Oct.2003, pp. 417–421.
[7] FLIR Systems, Inc. *Ladybug Cameras*. Accessed: May 1, 2020. [Online]. Available: https://www.flir.es/products/ladybug5plus/?model=LD5P-U3-51S5C-R
[8] S. Nayar, "Omnidirectional video camera," in *Proc. DARPA Image Understand. Workshop*, 1997, pp. 235–241.
[9] S. Baker and S. K. Nayar, "A theory of catadioptric image formation," in *Proc. 6th Int. Conf. Comput. Vis.*, 1998, pp. 35–42.
[10] J. Gaspar, N. Winters, and J. Santos-Victor, "Vision-based navigation and environmental representations with an omnidirectional camera," *IEEE Trans. Robot. Autom.*, vol. 16, no. 6, pp. 890–898, Jun. 2000.
[11] K. Okuyama, T. Kawasaki, and V. Kroumov, "Localization and position correction for mobile robot using artificial visual landmarks," in *Proc. Int. Conf. Adv. Mech. Syst. (ICAMechS)*, Aug. 2011, pp. 414–418.
[12] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, nos. 1–2, pp. 43–72, Nov. 2005, doi: 10.1007/s11263-005-3848-x.
[14] F. Wang, C. Liang, C. Ru, and H. Cheng, "An improved point cloud descriptor for vision based robotic grasping system," *Sensors*, vol. 19, no. 10, p. 2225, 2019.
[15] L. Payá, A. Peidró, F. Amorós, D. Valiente, and O. Reinoso, "Modeling environments hierarchically with omnidirectional imaging and global-appearance descriptors," *Remote Sens.*, vol. 10, no. 4, p. 522, Mar. 2018.
[16] T. Guan, Y. Fan, L. Duan, and J. Yu, "On-device mobile visual location recognition by using panoramic images and compressed sensing based visual descriptors," *PLoS ONE*, vol. 9, no. 6, 2014, Art. no. e98806.
[17] Z. Hu, B. Li, and Y. Hu, "Fast sign recognition with weighted hybrid K-Nearest neighbors based on holistic features from local feature descriptors," *J. Comput. Civil Eng.*, vol. 31, no. 5, Sep. 2017, Art. no. 04017034.
[18] L. Payá, O. Reinoso, Y. Berenguer, and D. Úbeda, "Using omnidirectional vision to create a model of the environment: A comparative evaluation of global-appearance descriptors," *J. Sensors*, vol. 2016, pp. 1–21, Mar. 2016.
[19] S. Cebollada, L. Payá, W. Mayol, and O. Reinoso, "Evaluation of clustering methods in compression of topological models and visual place recognition using global appearance descriptors," *Appl. Sci.*, vol. 9, no. 3, p. 377, Jan. 2019.

[20] Y. Li, Z. Hu, G. Huang, Z. Li, and M. A. Sotelo, "Image sequence matching using both holistic and local features for loop closure detection," *IEEE Access*, vol. 5, pp. 13835–13846, 2017.
[21] M. Horst and R. Möller, "Visual place recognition for autonomous mobile robots," *Robotics*, vol. 6, no. 2, p. 9, 2017.
[22] A.-D. Doan, Y. Latif, T.-J. Chin, Y. Liu, S. F. Ch'ng, T.-T. Do, and I. Reid, "Visual localization under appearance change: A filtering approach," in *Proc. Digit. Image Computing: Techn. Appl. (DICTA)*, Dec. 2019, pp. 1–8.
[23] F. Amorós, L. Payá, J. M. Marín, and O. Reinoso, "Trajectory estimation and optimization through loop closure detection, using omnidirectional imaging and global-appearance descriptors," *Expert Syst. Appl.*, vol. 102, pp. 273–290, Jul. 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417418301313
[24] E. Menegatti, T. Maeda, and H. Ishiguro, "Image-based memory for robot navigation using properties of omnidirectional images," *Robot. Auto. Syst.*, vol. 47, no. 4, pp. 251–267, Jul. 2004.
[25] S. Cebollada, L. Paya, V. Roman, and O. Reinoso, "Hierarchical localization in topological models under varying illumination using holistic visual descriptors," *IEEE Access*, vol. 7, pp. 49580–49595, 2019.
[26] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," in *Proc. Prog. Brain Res., Special Issue Vis. Perception*, vol. 155, 2006, pp. 23–26.
[27] F. Amorós, "Aplicación de la apariencia global de la información visual omnidireccional en color a tareas de navegación robótica en espacio $2\frac{1}{2}$D," Ph.D. dissertation, Dept. Ind. Eng. Syst. Automat., Univ. Miguel Hernández de Elche, Elche, Spain, Feb. 2014.
[28] F. Amorós, L. Payá, O. Reinoso, W. Mayol-Cuevas, and A. Calway, "Topological map building and path estimation using global-apperance image descriptors," in *Proc. Int. Conf. Informat. Control, Autom. Robot. (ICINCO)*, 2013, pp. 385–392.
[29] L. Payá, F. Amorós, L. Fernández, and O. Reinoso, "Performance of global-appearance descriptors in map building and localization using omnidirectional vision," *Sensors*, vol. 14, no. 2, pp. 3033–3064, Feb. 2014.
[30] C.-K. Chang, C. Siagian, and L. Itti, "Mobile robot vision navigation and localization using gist and saliency," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2010, pp. 4147–4154.
[31] A. J. Briggs, C. Detweiler, P. C. Mullen, and D. Scharstein, "Scale-space features in 1D omnidirectional images," in *Proc. 5th Workshop Omnidirectional Vis. (Omnivis)*, 2004, pp. 115–126.
[32] A. J. Briggs, Y. Li, and D. Scharstein, "Feature matching across 1D panoramas," in *Proc. 6th Workshop Omnidirectional Vis. (Omnivis)*, 2005, pp. 1–8.
[33] A. J. Briggs, C. Detweiler, Y. Li, P. C. Mullen, and D. Scharstein, "Matching scale-space features in 1D panoramas," *Comput. Vis. Image Understand.*, vol. 103, no. 3, pp. 184–195, Sep. 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1077314206000683
[34] H. Ishiguro and S. Tsuji, "Image-based memory of environment," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, vol. 2, Nov. 1996, pp. 634–639.
[35] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. Cognit. Neurosci.*, vol. 3, no. 1, pp. 71–86, Jan. 1991, doi: 10.1162/jocn.1991.3.1.71.
[36] M. Jogan and A. Leonardis, "Robust localization using eigenspace of spinning-images," in *Proc. IEEE Workshop Omnidirectional Vis.*, Jun. 2000, pp. 37–44.
[37] M. Jogan and A. Leonardis, "Robust localization using an omnidirectional appearance-based subspace model of environment," *Robot. Auto. Syst.*, vol. 45, no. 1, pp. 51–72, Oct. 2003.
[38] L. Payá, L. Fernández, A. Gil, and O. Reinoso, "Map building and Monte Carlo localization using global appearance of omnidirectional images," *Sensors*, vol. 10, no. 12, pp. 11468–11497, 2010. [Online]. Available: http://www.mdpi.com/1424-8220/10/12/11468
[39] F. Amorós, L. Payá, O. Reinoso, L. Fernández, and J. M. Marín, "Visual map building and localization with an appearance-based approach—Comparisons of techniques to extract information of panoramic images," in *Proc. 7th Int. Conf. Informat., Control, Autom. Robot. (ICINCO)*. Funchal, Portugal: SciTePress, 2010, pp. 423–426.
[40] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05)*, vol. 2, Jun. 2005, pp. 886–893.
[41] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," in *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[42] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.
[43] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[44] E. Hering, *Outlines of a Theory of the Light Sense*. Cambridge, MA, USA: Harvard Univ. Press, 1964.
[45] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Trans. Robot.*, vol. 25, no. 4, pp. 861–873, Aug. 2009.
[46] S. Sablak and T. E. Boult, "Multilevel color histogram representation of color images by peaks for omni-camera," in *Proc. SIP*, 1999, pp. 159–165.
[47] P. S. Suhasini, K. S. R. Krishna, and I. V. M. Krishna, "Combining SIFT and invariant color histogram in HSV space for deformation and viewpoint invariant image retrieval," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res. (ICCIC)*, Dec. 2012, pp. 1–4.
[48] C. Junhua and L. Jing, "Research on color image classification based on HSV color space," in *Proc. 2nd Int. Conf. Instrum., Meas., Comput., Commun. Control (IMCCC)*, Dec. 2012, pp. 944–947.
[49] ARVC (Automation, Robotics and Computer Vision Group). *Miguel Hernández University. Omnidirectional Images Database*. [Online]. Available: http://arvc.umh.es/db/images/quorumv/
[50] D. M. W. Powers, "Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation," School Inform. Eng., Flinders Univ., Adelaide, SA, Australia, Tech. Rep. SIE-07-001, 2007.
[51] L. Fernández, L. Payá, O. Reinoso, L. M. Jiménez, and M. Ballesta, "A study of visual descriptors for outdoor navigation using Google street view images," *J. Sensors*, vol. 2016, pp. 1–12, Dec. 2016.

FRANCISCO AMORÓS received the B.S. degree in industrial engineering from Miguel Hernández University, Elche, Spain, in 2009, the M.S. degree in industrial and telecommunication technologies, in 2010, and the Ph.D. degree in industrial technologies, in 2014. From 2010 to 2014, he was awarded with a Ph.D. Scholarship from the Generalitat Valenciana Government, Spain. Since 2014, he has been an Adjunct Professor with the Department of Electronics, Automation, and Communications, Comillas Pontificial University. He is also a Collaborator Researcher with Miguel Hernández University. His research interests include robotic navigation, omnidirectional vision, and image processing.

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

WALTERIO MAYOL-CUEVAS (Member, IEEE) received the B.Sc. degree from the National University of Mexico and the Ph.D. degree from the University of Oxford. He is a member of the Department of Computer Science, University of Bristol. His research with students and collaborators proposed some of the earliest versions of visual simultaneous localization and mapping (SLAM) and its applications to robotics and augmented reality. These include flagship humanoid robots and early commercial applications of visual mapping for wearable computing. His most recent works include novel concepts of human–robot interaction, fast computer vision methods for scene understanding, and algorithms for novel visual sensors and machine learning applications to assess its skill. He is the General Co-Chair of BMVC 2013 and the General Chair of the IEEE ISMAR 2016.

LUIS MIGUEL JIMÉNEZ received the degree in industrial engineering from the Polytechnic University of Madrid and the Ph.D. degree in robotics and automation from Miguel Hernández University, Elche, Spain. He is an Associate Professor with Miguel Hernández University, in the areas of systems engineering and automation. His research interests are focused in the fields of automation, robotics, and computer vision with the ARVC Research Group, Miguel Hernández University.

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