Line Image Signature for Scene Understanding with a Wearable Vision System

Alejandro Rituerto
DIIS - I3A,
University of Zaragoza, Spain
arituerto@unizar.es

Ana C. Murillo
DIIS - I3A,
University of Zaragoza, Spain
acm@unizar.es

J.J. Guerrero
DIIS - I3A,
University of Zaragoza, Spain
jguerrer@unizar.es

ABSTRACT

Wearable computer vision systems provide plenty of opportunities to develop human assistive devices. This work contributes on visual scene understanding techniques using a helmet-mounted omnidirectional vision system. The goal is to extract semantic information of the environment, such as the type of environment being traversed or the basic 3D layout of the place, to build assistive navigation systems. We propose a novel line-based image global descriptor that encloses the structure of the scene observed. This descriptor is designed with omnidirectional imagery in mind, where observed lines are longer than in conventional images. Our experiments show that the proposed descriptor can be used for indoor scene recognition comparing its results to state-of-the-art global descriptors. Besides, we demonstrate additional advantages of particular interest for wearable vision systems: higher robustness to rotation, compactness, and easier integration with other scene understanding steps.

Author Keywords

Wearable Omnidirectional vision system; Scene understanding; Visual assistance

ACM Classification Keywords

I.2.10. Artificial Intelligence: Vision and Scene Understanding

INTRODUCTION

The growing interest and developments on wearable computer vision systems are facilitating new systems and technologies for human assistance. Our goal is to provide a wearable indoor navigation assistance system with semantic information about the environment traversed by the user. In particular, this work is focused on the problem of scene understanding. The prototype developed consists of a helmet-mounted omnidirectional camera (see Fig. 1) aimed for indoor navigation assistance. Straight segments play an important role to understand the content of images from man made environments (see Fig. 2). People can easily guess the 3D structure of a scene represented by line sketches. Line and contour cues have been extensively used to analyze images since they provide very useful information. Contours occur as boundaries of objects, helping to detect them, or as frontiers between surfaces, encoding the structure of the scenes. Analyzing contours in the images have been shown useful for many different tasks, such as object recognition [3], 3D scene reconstruction [9] or image registration [16].

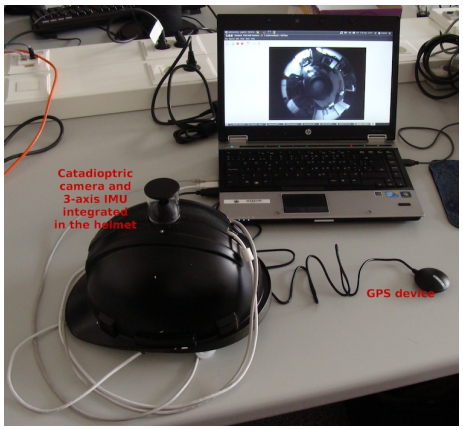


Figure 1. Wearable Omnidirectional camera system. The omnidirectional camera is mounted in a helmet. The system includes an IMU and a GPS device, but they are not used in this work.

We propose a novel line-based scene descriptor which is obtained as follows: scene lines are extracted from the omnidirectional images and classified in the three scene dominant directions; then, the descriptor is built as a histogram that encloses the distribution of these lines at different image regions. We use a catadioptric vision system (shown in Fig. 1) able to capture a 360^o field of view, therefore observed lines have the advantage of being longer that in conventional images. However, it presents highly distorted images, making line detection more difficult. Although this work is demonstrated in catadioptric systems, it can be easily extended to other omnidirectional vision systems.

Our experimental validation shows a detailed analysis of the parameters of the descriptor computation and demonstrate its performance for indoor scene recognition on a realistic dataset[1]. The place recognition capabilities of the proposed

SenseCam 2013 San Diego, USA

[1]Wearable OmniDirectional Vision System Dataset `http://robots.unizar.es/omnicam/`

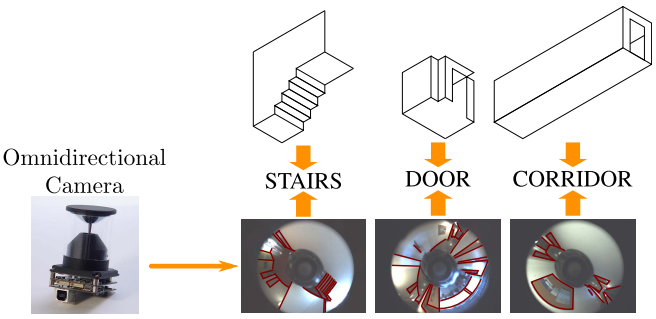


Figure 2. Top row shows line sketches representing different types of man made scenes easily recognized by a person. Bottom row shows lines extracted in catadioptric images of the same kind of places. Although not so intuitive for our sight, image lines (red) extracted in the omnidirectional images are as informative as the top row sketches to identify the type of scene represented.

descriptor are comparable to state of the art global image descriptors for this task, while it is shown to be more compact and presents higher rotation invariance (to rotation around the vertical axis of the camera). These additional advantages are important when working with a wearable system, due to the constant and heterogeneous movements done by a person. Besides, the proposed description method extracts and processes scene lines which are also required for following scene analysis steps, such as 3D layout reconstruction.

RELATED WORK

Different types of contour based image features have been used in many computer vision applications since they provide very distinctive information.

For example, one of the applications where line cues have shown great potential for 3D scene understanding from a single image. Straight lines are highly present in man made environments, in particular, parallel lines aligned with the main directions of the scene (Manhattan World assumption). Based on these cues, authors of [9] presents a method to extract the spatial layout of a room even with cluttered boundaries. The approach from [18] proposes an improvement of the performance of state-of-the-art methods for spatial layout computation by decomposing the potentials used in previous literature into more computationally tractable pair-wise potentials. We also find approaches specific for omnidirectional vision [14], that extract the spatial layout of indoor scenes from a single image.

Lines and boundaries have been also used for shape and object recognition tasks. Authors in [3] presented the shape context, which stores the relation between a contour point and the rest of the contour points of the shape. We also find works that use similarity measures defined for sets of connected contour segments for object recognition [7]. Line sketches were also shown to work well as models for object recognition in [6]. Other applications of image lines include recovering the rotation between frames for visual based localization [10] or their use for image retrieval [16].

Working with lines presents difficulties to obtain correspondences between images, usually because of the low accuracy or robustness of the line tip detection. However, lines present advantages for tasks that need to deal with extreme illumination changes or low textured environments [19], outperforming local point feature based methods for these settings [11]. We find approaches that propose to use straight line segments as local image features, describing them with different statistics around the edges. Many of these works make use of geometric constraints to obtain more robust matching results, e.g., homographies [17] or epipolar constraints [1]. Recent works have proposed more sophisticated line-based local descriptors, such as the Line Signature [19] that outperforms point based local features matching low textured images. MDSL descriptor [20] is another line-based local descriptor, which is built for each detected line segment. It is shown to be highly distinctive and robust to image rotation, illumination and viewpoint change.

Closer to our work, other approaches try to encode the image information with a line-based global descriptor [10]. Here, the authors propose the Line Histogram, which represents angles and lengths of all the boundaries of an image in a histogram. Our approach also creates a line-based global descriptor, but it captures the distribution of the scene lines in the omnidirectional image.

We have chosen global descriptors because they have shown good compromise between precision and computational cost for general scene recognition problems. In [13] a global Gist descriptor is presented for scene recognition in real world scenes. Authors in [5] present the Histogram of Oriented Gradients (HOG) which encodes the gradient orientations present at different image regions.

As already mentioned, this work is focused on omnidirectional cameras. Due to the wide FOV of this kind of systems, more lines of the scene appear in the image and they are longer that in conventional images. However, in the catadioptric vision system used these lines appear as conics in the image. We find works that have faced the use of lines in omnidirectional images for rotation estimation [2, 4]. Both papers propose a method to detect scene lines in omnidirectional images and compute their vanishing points. Estimating the vanishing points in omnidirectional images is typically robust and accurate, since these points are visible in the image. In this work we use the second work for line extraction and vanishing point computation.

LINE-BASED IMAGE SIGNATURE DESCRIPTOR

This section details the steps to obtain the proposed Line-based Image Signature (LIS) descriptor. First, the conics of the image, which actually correspond to straight lines of the scene, are extracted. Later, these conics are classified using the vanishing points information. Finally the descriptor is built as a set of histograms of the distribution of the classified contours in the image space.

Line extraction

The method used for the extraction of the scene lines in the catadioptric images was presented by Bermudez et al. in [4]. In this work, the authors describe a system to detect the conics projected in an omnidirectional image corresponding to

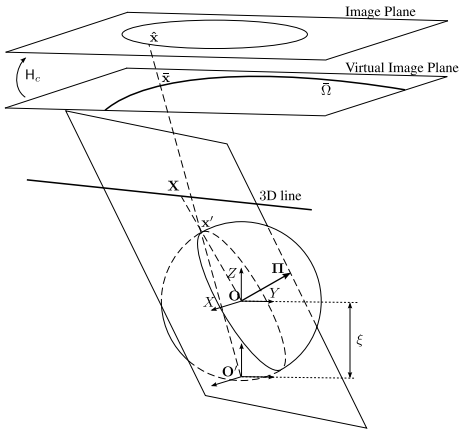
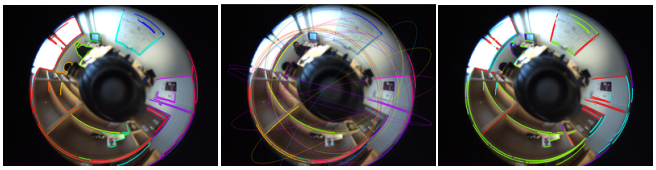


Figure 3. Projection of a 3D line and a point X in the line with the Spherical Camera Model.



(a) (b) (c)

Figure 4. (a) shows the edges extracted by the Canny algorithm and how they are grouped. Each color represents a group. (b) shows the conics extracted for certain groups of edges and (c) represents the edges once all the conics have been classified according to the vanishing points. Vertical vanishing point (red), Horizontal vanishing points (blue and green), and Non aligned conics (purple).

straight lines of the scene. The method requires the calibration of the camera and uses just two points to adjust a conic in the image.

The Spherical Camera Model [8] is used to model the omnidirectional camera projection. The projection of a 3D point in the image through this model is performed in three steps. First, the 3D point, $\mathbf{X}$, in the reference system of the camera, is projected into a unitary sphere centered in the effective viewpoint $\mathbf{O}$. The resulting point, $\mathbf{x}'$, is reprojected into the virtual image plane through $\mathbf{O}'$, whose distance to $\mathbf{O}$ is ξ. The relation between the virtual image plane and the real image is a collineation: H_c. The process is shown in Fig. 3.

In central catadioptric cameras, the projection of a 3D straight line results in a conic in the image. A 3D line, l, defines a plane, $\mathbf{\Pi}$, together with the effective point of the omnidirectional camera, $\mathbf{O}$. Given this plane $\mathbf{\Pi}$, the equation of the conic projected in the virtual image plane is

$$\bar{\Omega} = \begin{pmatrix} n_x^2(1-\xi^2) - n_z^2\xi^2 & n_x n_y(1-\xi^2) & n_x n_z \\ n_x n_y(1-\xi^2) & n_y^2(1-\xi^2) - n_z^2\xi^2 & n_y n_z \\ n_x n_z & n_y n_z & n_z^2 \end{pmatrix}, \tag{1}$$

where n_x, n_y and n_z are the components of the normal of the plane, $\mathbf{\Pi} = (n_x, n_y, n_z)^T$.

A point, $\bar{\mathbf{x}}$, is part of the conic if $\bar{\mathbf{x}}^T\bar{\Omega}\bar{\mathbf{x}} = 0$. The relation between the point coordinates and the plane formed by the

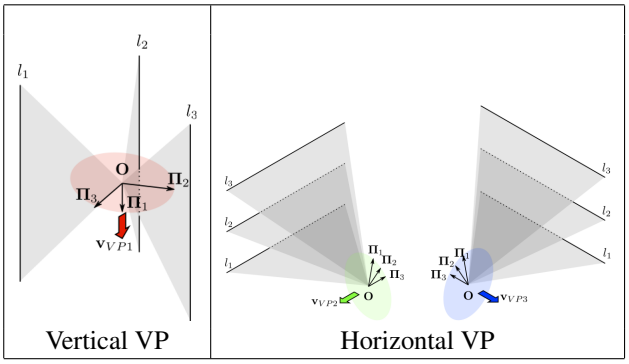


Figure 5. Sample sets of parallel lines and the corresponding VP directions. l_i denotes the line i and $\mathbf{\Pi}_i$ the normal of the plane created with O and represented by a gray surface. The normals of parallel lines are coplanar and perpendicular to the VP direction. The colored circles show the plane formed by the normals of parallel lines. These planes are perpendicular to the VP directions. Vertical VP, $\mathbf{v}_{VP1}$, (red), Horizontal VP, $\mathbf{v}_{VP2}$ and $\mathbf{v}_{VP3}$, (blue and green).

3D line is

$$\alpha = -\frac{\bar{z}}{1-\xi^2} \pm \frac{\xi}{1+\xi^2}\sqrt{\bar{z}^2 + (\bar{x}^2+\bar{y}^2)(1-\xi^2)} \tag{2}$$

where $\alpha = \frac{n_x\bar{x}+n_y\bar{y}}{n_z}$.

Given two points in the virtual image plane, $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, the image conic including both being the projection of a 3D line, can be computed solving the next system

$$\begin{pmatrix} \bar{x}_1 & \bar{y}_1 & -\alpha_1 \\ \bar{x}_2 & \bar{y}_2 & -\alpha_2 \end{pmatrix} \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{3}$$

Using this relation, the process of the line extraction is the following. First, the Canny algorithm is used to detect the edges that appear in the image. Connected edges are grouped together in boundaries. For each boundary a two point RANSAC is run in order to detect the conics formed by the edges. Process repeats till the number of edges in the boundary falls below a threshold, and no more conics are adjusted. Fig. 4 shows plots of different steps of the detection process.

Classification according to vanishing points

Once all the image conics of the scene have been detected, we classify them according to the vanishing points. The vanishing points (VP) are the image points where parallel lines intersect. In man made environments, we find three main vanishing points, which correspond to the vertical lines and two sets of horizontal lines.

The vanishing points lay in the infinite, so they are defined by a direction, $\mathbf{v}_{VP}$. As showed in Fig. 5 all the normals of the planes created by parallel lines and $\mathbf{O}$ are coplanar. They are also perpendicular to the corresponding direction of the VP where they intersect, $\mathbf{v}_{VPk}$. Being l_i, l_j and l_m 3 parallel lines intersecting in the kth VP, and $\mathbf{\Pi}_i$, $\mathbf{\Pi}_j$ ad $\mathbf{\Pi}_m$ their corresponding planes, then

$$\begin{aligned} (\mathbf{\Pi}_i \times \mathbf{\Pi}_j) \cdot \mathbf{\Pi}_k = 0 \\ (\mathbf{\Pi}_i \times \mathbf{\Pi}_j) \cdot \mathbf{v}_{VPk} = 1 \end{aligned} \tag{4}$$

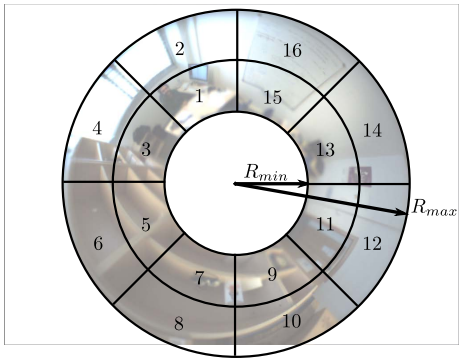


Figure 6. Image tessellation to incorporate in the descriptor the spatial distribution of scene lines (in this case $n = 2$).

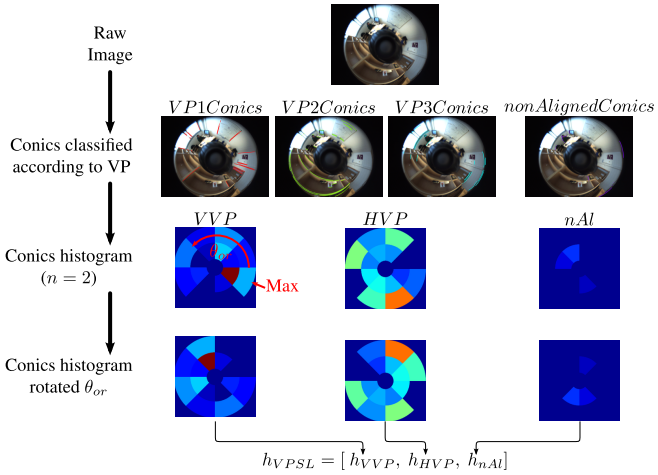


Figure 7. Steps to build the descriptor (from top to bottom). From the raw image we obtain the aligned conics, conics are then classified according to the VP of the scene and the histograms are built. Third row shows the sample histograms in polar coordinates (the color of each bin is related with the number of edges in that bin). Last row shows the final histogram, once the rotation invariance orientation has been obtained.

These properties can be used to group the conics according to the vanishing points. We start with all the detected conics, each one represented by a $(\mathbf{\Omega}_i, \mathbf{\Pi}_i)$ pair. We assume an approximate vertical position of the camera, so we can predefine prior directions for the vanishing points: $\mathbf{v}_{VP1}$ is vertical and $\mathbf{v}_{VP2}$ and $\mathbf{v}_{VP3}$ are parallel to the image plane. We use a robust estimation algorithm (RANSAC) to adjust this estimation and obtain one group of conics corresponding to each of these main directions. In each RANSAC iteration two conics are randomly selected to create a group hypothesis. If these conics are parallel and aligned with the corresponding VP, $|(\mathbf{\Pi}_1 \times \mathbf{\Pi}_2) \cdot \mathbf{v}_{VPk}| \geq 1 - thP$, we check how many of the rest of the conics are parallel to them and vote for that hypothesis, $(\mathbf{\Pi}_1 \times \mathbf{\Pi}_2) \cdot \mathbf{\Pi}_j \leq thC$. The most voted hypothesis is chosen as one of the groups. Once the groups for the three VP have been obtained, the remaining conics are grouped as $nonAlignedConics$.

The result of this process is useful not only for globally describe the image. The information of VP aligned conics can be used to perform other scene understanding tasks such as 3D analysis of the scene under certain assumptions.

Building the descriptor

Once all the detected conics have been classified according to their VP, we can build the proposed LIS descriptor. To build the descriptor, the image space is discretized in a polar grid. The image is split in $4 \times n$ angular sections and each angular section is split into n radial sections. Modifying n we can adjust the size and resolution of the descriptor.

Due to the self-reflection some parts of the camera and the mirror support appear in the image. In our system, these image parts are constrained by the minimum and maximum radius, R_{min} and R_{max} respectively. With this discretization, each histogram will be compound of $4 \times n \times n$ bins. Figure 6 shows the grid and the bins of the discretization for the simplest case $n = 2$.

We create histograms for the vertically aligned conics, h_{VVP}, for the horizontally aligned conics, h_{HVP} and for the non aligned conics, h_{nAl}. The value of bin i of each histogram is

$$\begin{aligned} h_{VVP\,i} &= 100\frac{\#\text{ Vertically aligned edges in bin } i}{\#\text{ Total edges}} \\ h_{HVP\,i} &= 100\frac{\#\text{ Horizontally aligned edges in bin } i}{\#\text{ Total edges}} \\ h_{nAl\,i} &= 100\frac{\#\text{ non aligned edges in bin } i}{\#\text{ Total edges}} \end{aligned} \tag{5}$$

where $i \in [1..4 \times n \times n]$. Histogram h_{VVP} is compound by the conics included in $VP1Conics$, h_{HVP} by the conics in $VP2Conics$ and $VP3Conics$, and the h_{nAl} histogram by $nonAlignedConics$. The total number of edges, $\#$ Total edges, correspond to the sum on all the edges of all the detected conics.

When grouping the conics of a scene, the two possible horizontal VP could be misclassified due to a different orientation of the camera in the same scene. In order to avoid this we join the conics aligned with the horizontal vanishing points, $VP2$ and $VP3$, in the same histogram, h_{HVP}. The final descriptor, h_{LIS}, is composed of the three histograms organized as follows:

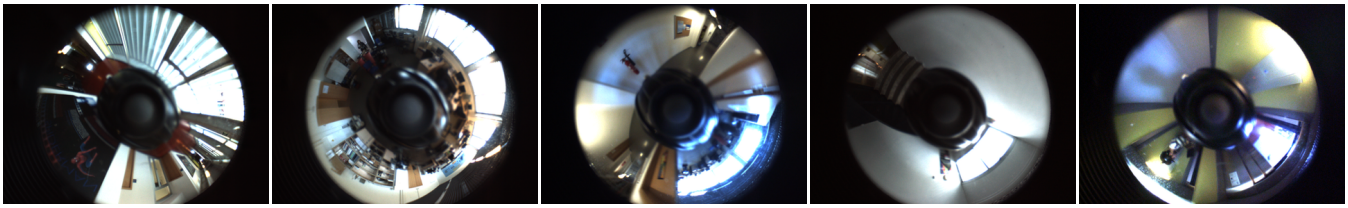
$$h_{LIS} = [h_{VVP},\ h_{HVP},\ h_{nAl}]. \tag{6}$$

For omnidirectional cameras, rotation invariance is an important property to be able to recognize a scene when facing it with different direction of travel. To achieve rotation invariance, we have defined a common reference for all the images. We set the reference for each image to the angular segment where most of the vertical line edges lie. This segment gives us the orientation angle θ_{or}. Fig. 7 represents the described process from top to bottom.

Image similarity using the LIS descriptor

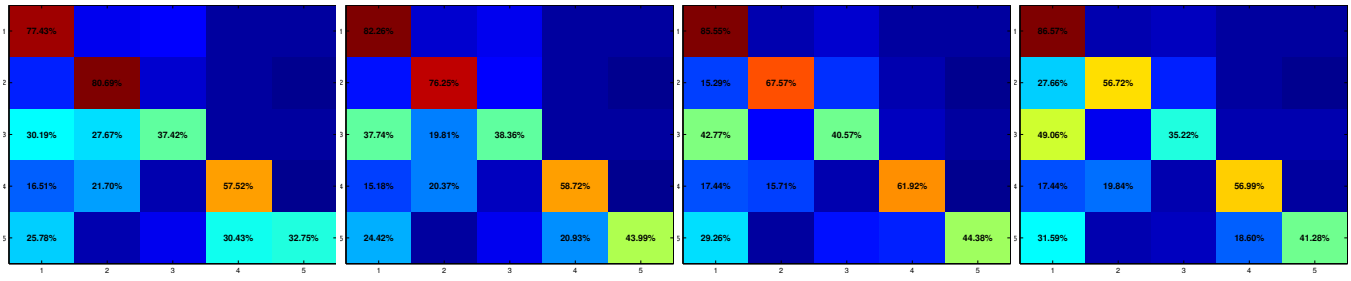
The distance between two of LIS descriptors can be simply computed as the absolute distance between histograms. Given the LIS descriptor of two images, h_1 and h_2, the distance d between them is

$$d = \sum_{i=1}^{3\cdot 4\cdot n\cdot n} \|h_{1i} - h_{2i}\|. \tag{7}$$



Class 1 - Corridor | Class 2 - Room | Class 3 - Door/Jamb | Class 4 - Stairs | Class 5 - Elevator

Figure 8. Examples of the images included in the dataset for each considered category.



$n = 2$ $n = 4$ $n = 8$ $n = 16$

Figure 9. Confusion matrices for different image space discretization (n). Each row shows how many tests of that class were classified as any of the possible classes (1: Corridor, 2: Rooms, 3: Doors or Jambs, 4: Stairs, 5: Elevator). Only numeric values above 15% are shown. Color goes from Dark blue for 0% to Dark red for 100%.

	Train	Test
Class 1	61%	70%
Class 2	19%	12%
Class 3	9%	4%
Class 4	9%	10%
Class 5	2%	4%

Table 1. Percentage of frames of each class in each sequence.

EXPERIMENTS

This section analyzes the properties of the proposed scene signature through experiments with real images, acquired with a calibrated wearable catadioptric vision system at 1024×768 resolution. The dataset is detailed in the following subsections.

LIS image tessellation analysis

This experiment evaluates the performance of our proposed descriptor for image categorization with different tessellation values. We used the OmniCam dataset presented in [15], and used for metric and topological indoor navigation. The dataset consists of two different sequences of images acquired in the same environment. To evaluate the robustness of the place categorization, the two sets were acquired at different times (several months between acquisitions) and covering different trajectories, i.e., some areas covered by the test sequence do not appear in the training set. The sequence used as training set in this work includes about 12200 catadioptric image frames, and the test set includes about 7300 frames. All the images were manually labeled with the corresponding type of indoor area to set the labels ground truth. The 5 classes used to describe the areas of our indoor environment are shown in Fig. 8 together with some examples of images in the dataset. The percentage of frames of each class for the training set is shown in Table 1. It can be observed that some classes have many more examples than others, e.g., 61% corridor images and 2% of elevator images, due to typical configuration of indoor environments, since any user spends more time traversing corridors than in the elevators, and the whole sequences have been used.

The experiment consists on assigning a scene class label to each test image, according to the nearest neighbour found among the training images. We choose a nearest neighbour based approach due its simplicity and because other standard classification frameworks such as SVM or boosting based approaches typically require more training data. Probably the use of more complex learning techniques will raise the performance with any of the descriptors evaluated. Test images are compared with all the images of the training set, and the query is labeled with same class of the train image with the lowest descriptor distance, computed as described in (7). We run this experiment with different configurations of our approach. Fig. 9 shows the confusion matrices of the classification using different image tessellation: $n = 2, 4, 8$ and 16. Higher values of n mean larger (descriptor size: $3 \times 4 \times n \times n$) and higher resolution descriptors. Looking at the results we can see how the performance is not homogeneous for all the classes. Corridors (Class 1) are recognized better in all the experiments, probably because they have well-defined vanishing points and lines. The worst performance corresponds to Doors/Jambs (Class 3). When traversing a door or a jamb, large part of the omnidirectional image contains areas of the environment before and after that door, so confusion using the description of the whole omnidirectional image seems reasonable.

Parameter n is directly related with the size and the resolution of the descriptor. In general, increasing n improves the performance, however, there is a point where higher n values

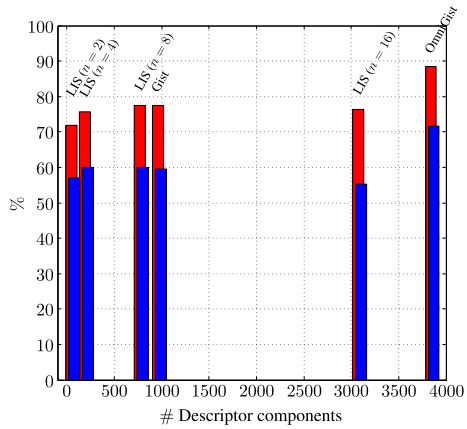


Figure 10. Average precision (red) and Average class precision (blue) for the descriptors tested as function of the descriptor size.

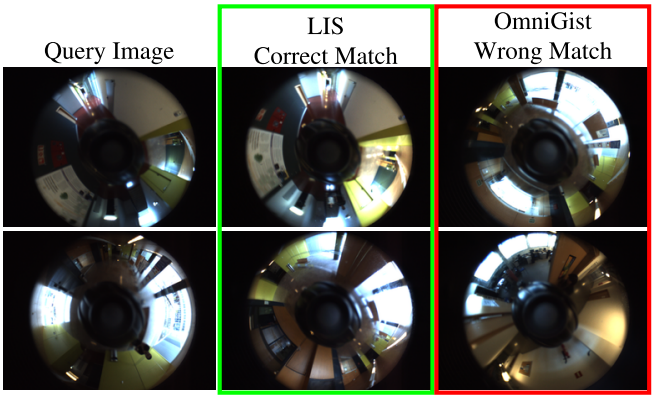


Figure 11. Examples of images where LIS descriptor performs better than the other descriptors. The first column shows the query images, second and third columns show the nearest train image selected by LIS and OmniGist respectively. Rotation between the query image and the correct match selected with LIS can be notice in the figure.

does not improve the result, even performance is decreased. For our settings the best value is $n = 8$. We should note the behavior of the recognition for the Rooms (Class 2). For this class, an increment of n produces, in all the cases, worst results. This class groups many kind of spaces (halls, laboratories, offices) with different appearance but similar structure. In this case an increment of the resolution of the descriptor reduces the performance because it starts enclosing too subtle details that are not common to all elements in the class.

Comparison with state-of-the-art descriptors

This subsection shows the performance of our proposal compared to other global descriptors used for scene recognition in omnidirectional images. We compare three global descriptors: our proposed LIS approach, the Gist descriptor [13] using the code provided by the authors, and the adaptation of this descriptor to omnidirectional images [12].

The average precision and average class precision for all the descriptors are shown in Fig. 10. The average precision (red) and the average class precision (blue) are plot as function of the number of components of each descriptor: $3 \times 4 \times n \times n$ for LIS, 920 for Gist and 3480 for OmniGist. Our proposed descriptor achieves a precisions between 72% and 78% depending on the discretization, Gist gets a precision of 83% and OmniGist 88%, and similarly for the average class precision. The precision of our approach is slightly lower for the used dataset, but the shorter length descriptor is an important advantage for navigation applications where memory and computational power can be a limitation. Fig. 11 shows some examples where LIS ($n = 8$) classify correctly the category of the query image while the OmniGist fails.

The dataset used was captured while traversing a typical indoor environment, therefore for most of the cases train and test images are related by multiples of 90^o rotation. Therefore, the more robust rotation invariance of our descriptor cannot be observed using only that dataset. Next experiment proves the higher rotation invariance of our approach.

Rotation invariance

Rotation invariance is an interesting property to achieve robust place recognition when working with omnidirectional images. It allows to have less training data (we do not need to have examples of every possible acquisition angle at each location) but still a robust modeling of the different classes. This is one important advantage of our proposed descriptor.

Next experiment analyzes the rotation invariance of our approach compared to other global descriptors. We get 36 images equally distributed along a 360^o camera rotation movement, around the vertical camera axis. The camera used for this experiment is different than the one used in the rest of experiments. It has been installed a goniometer for angle measurement and the images resolution is 1024×768. Fig. 12 presents a plot of the descriptor distance versus the angular difference between all images in the sequence and the first one. To be able to compare the distances of different descriptors, the distance values have been normalized using Standard Score normalization. Descriptor distances are shown as the difference with the mean of the distances, μ, in units of the standard deviation, σ.

For the standard Gist descriptor, we observe it is not rotation invariant The minimum distance appears in 0^o and 360^o, but grows continuously to the maximum value at rotations around 180^o. For the OmniGist descriptor, minimum values of the distance appear for angle differences multiple of 90^o, while maximum values appear for angles multiple of 45^o. Finally, for our approach the distance also varies with angle, the maximum distance occurs for rotations around 180^o, however we can observe local minimum points around rotations multiple of 45^o. This is due to the image space discretization used for the descriptor, $n = 2$, where each angular segment corresponds to an interval of 45^o. The variation of the distances, between images in the sequence and the reference image at 0^o, is much lower for our proposed method, showing higher invariance to rotations around the vertical axis.

Integration with spatial layout extraction steps

As already described, our goal is to integrate the proposed place recognition in our navigation assistance system. This

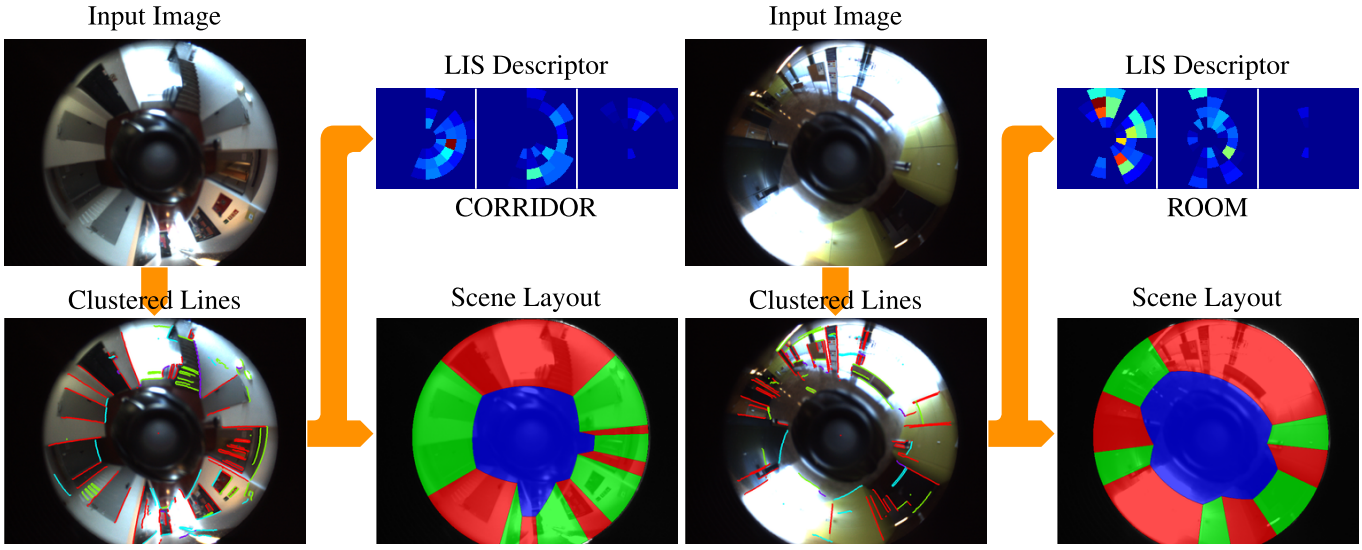


Figure 13. Two examples of the joint process of scene categorization and layout recovery for a corridor image (a) and a room image(b). First we extract and cluster the scene lines from the original image. With the line information we can compute the LIS descriptor ($n = 4$ shown) and get the kind of area being traversed. Besides, the layout recovery approach uses the same line information to get the scene planes of the scene, floor (blue) and walls (red and green) in this case.

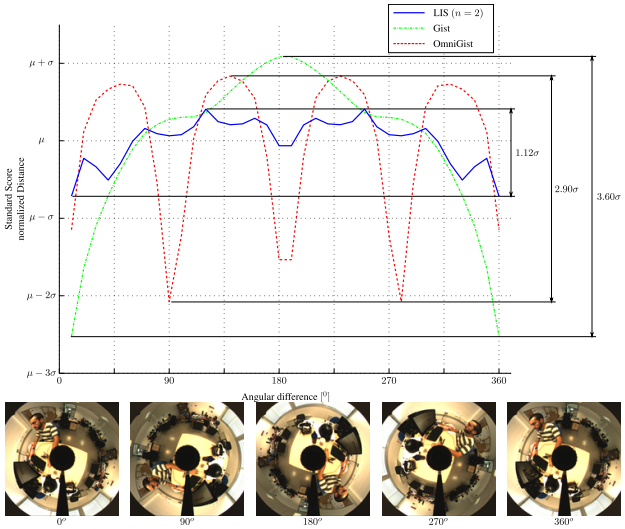


Figure 12. Rotation invariance analysis. Relation between the descriptor distance (vertical axis) and the angle difference (horizontal axis) among images of the same scene, acquired at the same location only with varying angles around the vertical axis. The proposed LIS descriptor presents much lower variation than other global descriptors.

work shows how to use of lines for scene recognition, but the same line information can be used for additional tasks such as 3D layout recovery. Using LIS global descriptor for place recognition, our system is able to detect when the kind of area traversed has changed and the category of the new place. The 3D scene analysis, that is computationally expensive, can then be run just when necessary, i.e., after a change of the type of area visited or for certain types of places.

As initial example of further exploitation of the lines detected we have run the code presented in [14] to detect the scene layout in omnidirectional images. This approach follows a heuristic to iteratively fit the wall-floor boundary, which follows the same steps and priors for any location. The same lines used to build the LIS descriptor are used to detect the scene layout. Fig. 13 shows the results of both tasks: the LIS histogram for scene categorization and the spatial layout of the image, where floor and walls are detected.

CONCLUSIONS

Scene understanding is an essential step towards many visual assistance applications that require semantic information of the environment. In this work we have presented a new line-based global descriptor for omnidirectional images that encloses the structure of the scene observed, by encoding the distribution of lines in the scene. We compute the conics in the image that correspond to lines of the environment and then we classify them according to the vanishing points. The descriptor is built as a histogram that captures how the different types of lines lay in the different parts of the image space. We have evaluated the system with catadioptric images but it is easily extensible to other panoramic systems.

We have run exhaustive experiments to analyze the properties of the proposed descriptor. Experiments for scene categorization show that the performance of our proposed descriptor is close to state-of-the-art global descriptors. Besides, we have shown our approach to have interesting advantages for the aimed application: small size, reducing the memory consumption and comparison time, and higher rotation invariance are interesting properties for person-mounted cameras. Additionally, preliminary results of integration with complementary line based scene understanding techniques are shown. Future work includes a robust integration of the scene categorization step with further scene analysis and guidance instructions steps.

ACKNOWLEDGMENTS

This work was supported by Spanish projects DPI2012-31781 and DGA T04-FSE and TAMA.

REFERENCES

1. Bay, H., Ferrari, V., and Van Gool, L. Wide-baseline stereo matching with line segments. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).* (2005).
2. Bazin, J.-C., Demonceaux, C., Vasseur, P., and Kweon, I. Rotation estimation and vanishing point extraction by omnidirectional vision in urban environment. *The International Journal of Robotics Research 31*, 1 (2012), 63–81.
3. Belongie, S., Malik, J., and Puzicha, J. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 4 (2002), 509–522.
4. Bermdez, J., Puig, L., and Guerrero, J. J. Hypercatadioptric line images for 3d orientation and image rectification. *Robotics and Autonomous Systems 60* (2012), 755–768.
5. Dalal, N., and Triggs, B. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2005).
6. Eitz, M., Hildebrand, K., Boubekeur, T., and Alexa, M. Sketch-based 3d shape retrieval. In *SIGGRAPH 2010: Talks* (2010).
7. Ferrari, V., Fevrier, L., Jurie, F., and Schmid, C. Groups of adjacent contour segments for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 1 (2008), 36–51.
8. Geyer, C., and Daniilidis, K. A unifying theory for central panoramic systems and practical applications. In *European Conference on Computer Vision (ECCV)* (2000), 445–461.
9. Hedau, V., Hoiem, D., and Forsyth, D. Recovering the spatial layout of cluttered rooms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2009).
10. Kosecká, J., and Zhang, W. Video compass. In *European Conference on Computer Vision (ECCV)* (2002).
11. Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60*, 2 (2004), 91–110.
12. Murillo, A. C., Campos, P., Koeck, J., and Guerrero, J. J. Gist vocabularies in omnidirectional images for appearance based mapping and localization. In *10th OMNIVIS, held with Robotics: Science and Systems (RSS)* (2010).
13. Oliva, A., and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision 42*, 3 (2001), 145–175.
14. Omedes, J., Lpez-Nicols, G., and Guerrero, J. *Omnidirectional Vision for Indoor Spatial Layout Recovery*, vol. Frontiers of Intelligent Autonomous Systems. Springer, 2013, 95–104.
15. Rituerto, A., Murillo, A. C., and Guerrero, J. J. Semantic labeling for indoor topological mapping using a wearable catadioptric system. *Robotics and Autonomous Systems* (2012).
16. Russell, B., Efros, A. A., Sivic, J., Freeman, B., and Zisserman, A. Segmenting scenes by matching image composites. In *Advances in Neural Information Processing Systems* (2009).
17. Sagüés, C., Murillo, A. C., Escudero, F., and Guerrero, J. J. From lines to epipoles through planes in two views. *Pattern Recognition 39*, 3 (2006), 384–393.
18. Schwing, A. G., Hazan, T., Pollefeys, M., and Urtasun, R. Efficient structured prediction with latent variables for general graphical models. In *International Conference on Machine Learning (ICML)* (2012).
19. Wang, L., Neumann, U., and You, S. Wide-baseline image matching using line signatures. In *IEEE International Conference on Computer Vision (ICCV)* (2009).
20. Wang, Z., Wu, F., and Hu, Z. Msld: A robust descriptor for line matching. *Pattern Recognition 42*, 5 (2009), 941–953.