LINE-BASED GLOBAL DESCRIPTOR FOR OMNIDIRECTIONAL VISION

Alejandro Rituerto, Ana C. Murillo, J. J. Guerrero

Instituto de Investigación en Ingeniería de Aragón (I3A), University of Zaragoza, Spain
{arituerto, acm, josechu.guerrero}@unizar.es

ABSTRACT

Scene understanding is a widely studied problem in computer vision. Many works approach this problem in indoor environments assuming constraints about the scene, such as the typical Manhattan World assumption. The goal of this work is to design and evaluate a global descriptor for indoor panoramic images that encloses information about the 3D structure. This descriptor is based on the detection of representative lines of the scene, which encode the scene structure. Our work focuses on omnidirectional imagery, where observed lines are longer than in conventional images and the whole scene is captured in a single image. Experiments using two public datasets analyze the performance of the descriptor for scene categorization. We also analyze the influence of different parameters and show sample results for a navigation assistance application.

Index Terms— Panoramic image global descriptor; Omnidirectional images; Scene categorization

1. INTRODUCTION

This work is focused on the problem of scene understanding on indoor environments, where lines are known to play an important role (see Fig. 1): 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 tasks such as object recognition [1], 3D scene reconstruction [2] or image registration [3].

Our proposed line-based scene descriptor is obtained as follows: scene lines are extracted from the images and classified according to the three scene vanishing directions; The descriptor is built as a histogram that encloses the distribution of these lines in the image space. The descriptor is intended for omnidirectional images, where the whole scene can be captured in just one image. Lines appearing in theses images are longer than in conventional images and the vanishing points appear in the image, however they present high distortion, making line detection more complicated.

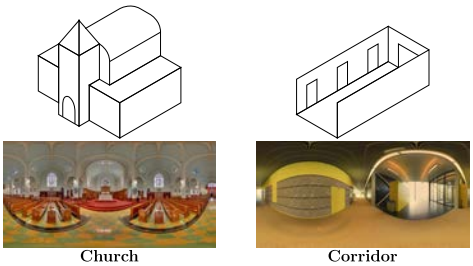


Fig. 1. Top row shows line sketches representing a church and a corridor. Bottom row shows panoramic images corresponding to these environments. The lines detected in the images are as representative of the kind of environment as the top row sketches.

The designed descriptor is compact and invariant to rotations around the vertical axis, important requirements when working with robots or autonomous systems. Besides, the proposed descriptor extracts and processes scene lines in a way that can be used for other scene analysis techniques, such as the 3D layout recovery. Experimental validation shows a detailed analysis of the parameters of the descriptor computation and demonstrates its performance for indoor scene categorization on a known public dataset [4], showing promising results.

2. RELATED WORK

Different types of contour based image features have been used in computer vision applications since they provide very distinctive information. One of the applications where line cues have shown great potential is in 3D scene understanding from a single image. Lines are present in man made environments. Under the Manhattan World assumption, lines are aligned with the dominant directions of the scene. Based on these cues, authors of [2] present a method to extract the spatial layout of a room even with cluttered boundaries. The approach in [5] achieves great performance by decomposing the potentials used in previous literature into more computationally tractable pair-wise potentials. Specific approach for omnidirectional vision [6] extracts the spatial layout of indoor scenes from a single image.

This work was supported by Spanish projects DPI2012−31781, DGA T04−FSE and TAMA

(a) (b) (c)

Fig. 2. Boundaries detection and classification. (a) Edges detected by the Canny algorithm grouped into connected boundaries. Each color represents a boundary. (b) Boundaries and corresponding to lines of the scene. (c) Lines classified according to the vanishing points: Vertical lines (red), Horizontal lines (blue and green), and non aligned lines (purple). (Best seen in color).

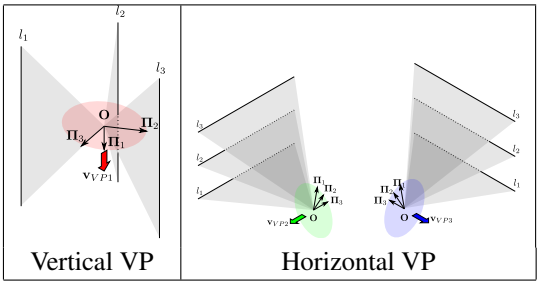


Fig. 3. Sample sets of parallel lines and the corresponding VP directions. l_i denotes the line i and $\mathbf{\Pi}_i$ the normal of the plane created with $\mathbf{O}$ and represented by a gray surface. The normals of parallel lines are coplanar and perpendicular to the VP direction. The colored circles 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).

Lines and boundaries have been also used for shape and object recognition tasks. The work in [1] present the shape context, which stores the relation between contour points. In [7] similarity measures for sets of connected contours are used for object recognition. Line sketches work as models for object recognition in [8]. Other applications include recovering the rotation between frames [9] or the use of lines for image retrieval [3].

Working with lines presents difficulties to obtain accurate correspondences between images, usually because of the low accuracy or robustness of the line detection. However, lines present advantages for tasks that need to deal with extreme illumination changes or low textured environments [10], outperforming local point feature based methods for these settings [11]. We find approaches that propose to use straight line segments as local image features. Many of these works use geometric constraints to obtain more robust matching results, e.g., homographies [12] or epipolar constraints [13]. Recent works propose more sophisticated line-based local descriptors, such as the Line Signature [10] that outperforms point based features for low textured images. MDSL descriptor [14] is built for each detected line segment and it is highly distinctive and robust to image rotation, illumination and viewpoint change.

Global descriptors have shown good compromise between precision and computational cost for general scene recognition problems. In [15] a global Gist descriptor is presented for scene recognition in real world scenes. This descriptor has been adapted to omnidirectional images in [16]. Work in [17] presents the Histogram of Oriented Gradients (HOG) which encodes the gradient orientations present at different image regions.

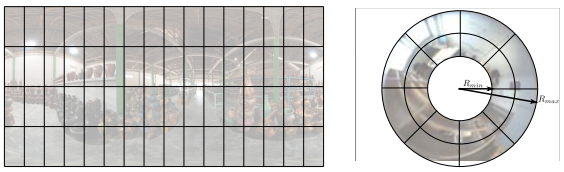
Closer to our approach, [9] proposed to encode the image information with a line-based global descriptor. Authors proposed the Line Histogram, which represents angles and lengths of the boundaries of an image in a histogram. Our approach also creates a line-based global descriptor, but it captures the distribution of the scene lines in the omnidirectional image.

3. LINE-BASED IMAGE SIGNATURE DESCRIPTOR

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

3.1. Omnidirectional line extraction

In omnidirectional images, lines of the environment are not projected as straight lines in the image. However, since the projection model is known, these 3D lines can be detected. The vision systems studied in this work are central systems, so the projection of the scene in the image is performed trough a single point $\mathbf{O}$ known as the camera center. When a straight line $\mathbf{l}$ is projected in the image trough this point, a plane $\mathbf{\Pi}$ is defined. To extract the lines in an image, the first step is to detect the boundaries in the image. This is done using a simple Canny edge detector and grouping the connected edges into boundaries.



(a) (b)

Fig. 4. Image tessellation used to compute the distribution of scene lines for (a) panoramic images ($n = 4$), and (b) catadioptric images ($n = 2$).

Equirectangular panoramas. In equirectangular images, coordinates in the image, relate linearly to pan and tilt angles in the real world. To extract the lines from the boundaries, a RANSAC algorithm is used. Given two edges of a boundary and the camera center, a plane can be computed. RANSAC algorithm looks for planes containing the most number of edges. These planes correspond to lines in the scene.

Catadioptric images. In this kind of systems, the boundaries corresponding to the projection of 3D lines are conics. We use the method presented in [18] for lines extraction in catadioptric images. This method requires the calibration of the camera and uses just two points to adjust a conic in the image.

3.2. Line classification according to vanishing points

Once all the lines of the scene have been detected in the image, 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: one vertical and two horizontal. The vanishing points lay in the infinite, so they are defined by a direction, $\mathbf{v}_{VP}$. As showed in Fig. 3 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, $\mathbf{v}_{VPk}$.

These properties can be used to group the boundaries according to the vanishing points. Each detected straight line is represented by a plane $\mathbf{\Pi}$, defined by the line $\mathbf{l}$ and the effective point of the vision system, $\mathbf{O}$. We use RANSAC again to adjust the vanishing directions and obtain one group of lines corresponding to each of these directions. Once the groups for the three VP have been obtained, the remaining boundaries are grouped as $nonAlignedBnds$.

The result of this process is useful not only for globally decribing the image. The information of VP aligned lines can be used to perform other scene understanding tasks such as the analysis of the 3D scene layout.

3.3. Building the descriptor

Once all the detected 3D lines have been classified according to their VP, we can build the proposed LIS descriptor. To build the descriptor, the image space is discretized. The width of the image is split in $4 \times n$ sections and each section is split into n height sections. With this discretization, each histogram will be compound of $4 \times n \times n$ bins. Figure 4 shows the image discretization for panoramic and catadioptric images.

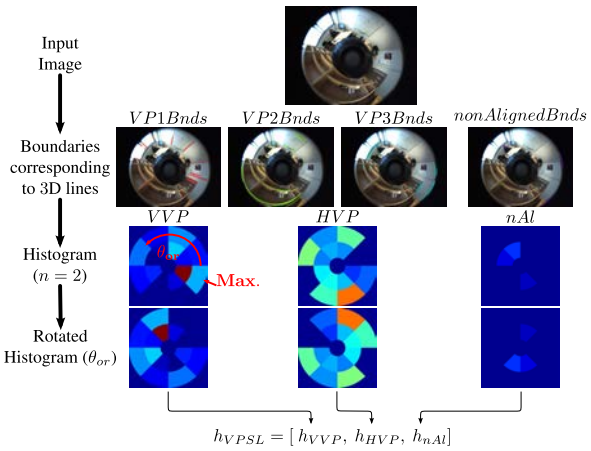


Fig. 5. Steps to build LIS using a catadioptric image, from top to bottom: from the raw image (first row), boundaries are extracted and classified according to the VP of the scene (second row); third row shows the example histograms in polar coordinates and last row shows the final histogram, after the rotation invariance step has been applied. The process is the same for panoramic images. (Best seen in color).

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

$$h_{VVP\ i} = 100 \frac{\#\text{ Vertically aligned edges in bin } i}{\#\text{ Total edges}} \quad (1)$$

$$h_{HVP\ i} = 100 \frac{\#\text{ Horizontally aligned edges in bin } i}{\#\text{ Total edges}} \quad (2)$$

$$h_{nAl\ i} = 100 \frac{\#\text{ non aligned edges in bin } i}{\#\text{ Total edges}} \quad (3)$$

where $i \in [1..4 \times n \times n]$.

Histogram h_{VVP} is compound by the boundaries included in $VP1Bnds$, h_{HVP} by the boundaries in $VP2Bnds$ and $VP3Bnds$, and the h_{nAl} histogram by $nonAlignedBnds$. The total number of edges, $\#$ Total edges, correspond to the sum on all the edges of all the detected boundaries. When grouping the boundaries of a scene, horizontal VP can be confused due to a different orientation of the camera in the same scene. In order to avoid this we join the boundaries 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:

$$h_{LIS} = [h_{VVP},\, h_{HVP},\, h_{nAl}]. \quad (4)$$

When using mobile systems with omnidirectional cameras, rotation invariance is an important property to be able to recognize the same scene when facing it

from different travel directions. 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. 5 represents the described process from top to bottom. Figure shows the process for a polar grid (catadioptric image), however, the same process is used for the panoramic images with a rectangular grid. Both systems capture the whole scene (360^o).

Comparing images with the LIS descriptor. To compare the LIS descriptors of two different images we simple use the L_1 norm between the histograms.

4. EXPERIMENTS

This section shows the performance of the LIS descriptor for scene categorization using a panoramic dataset and for navigation assistance using images from a catadioptric vision system.

4.1. Scene Categorization performance

For this scene categorization experiment, we use the publicly available SUN360 dataset, presented in [4]. It contains $67,583$ panoramas of 80 categories. The panoramas are labeled as indoor or outdoor scenes and further classified into more detailed categories. In this work just indoor images are used: $3,889$ images classified into 15 categories (cave, church, corridor, hotel room, living room, lobby atrium, museum, old building, restaurant, shop, showroom, subway station, theater, train interior and workshop).

The goal in scene categorization is to detect the type of environment where an image was acquired. Our experiment consists of classifying the test images into the correct class by comparing them to the reference set using our proposed descriptor. We randomly split the dataset into test and reference sets: 70% of the images are used for test, and the reference set consists of the remaining 30%. All the images in the test set are compared to the images in the reference set, and the class assigned is the class of the closest reference image (Nearest Neighbor classification).

Table 1 shows the Total and Average per class accuracy of the classification for different tessellations, and the descriptor size. Best total accuracy is achieved for $n = 8$, when the descriptor includes 768 components. It has to be noticed that the accuracy differences are small compared with the descriptor size, bigger descriptor mean higher resolutions but do not represent better performance. Best performance correspond to hotel room (46.79%) and subway station (38.42%), however, the descriptor accuracy falls for old building (8.13%) and shop (7.12%).

4.2. LIS for Navigation assistance

The proposed descriptor could also be used for navigation tasks as shown in [19]. In this experiment the

	Accuracy		Descriptor
	Total (%)	Average (%)	size
$n = 2$	18.81	16.69	48
$n = 4$	18.59	16.50	192
$n = 8$	19.10	16.92	768
$n = 16$	18.81	17.45	3072

Chance: 6.67%

Table 1. Total and average accuracy of the scene categorization for different image tessellations.

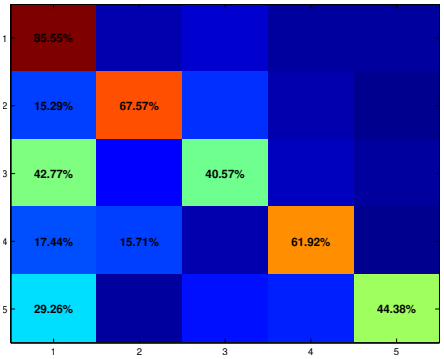


Fig. 6. Confusion matrix for catadioptric images ($n = 8$). Each row shows how many tests of that class were classified as any of the possible classes (1: Corridor, 2: Rooms, 3: Doors or Jambs, 4: Stairs, 5: Elevator). Only numeric values above 15% are shown. Color goes from Dark blue for 0% to Dark red for 100%. (Best seen in color).

LIS descriptor is used as part of a wearable navigation assistance system, where the images should be classified into 5 classes: Room, Corridor, Door, Stairs and Elevator. Fig. 6 shows the confusion matrix of scene categorization using a public dataset of indoor omnidirectional images (OmniCam dataset [20]). The accuracy values in this experiment are better than in the previous one, from 40% to 85%, but one should note that it is a less challenging problem, with less classes and images acquired in the same environment. So if we compare with the chance classification in each experiments, results in both cases are equally promising.

5. CONCLUSIONS

In this work we have presented a line-based global descriptor for omnidirectional and panoramic images that encloses the structure of the scene observed. The descriptor is built as a histogram capturing how the scene lines lay in the image space. A public dataset has been used to evaluate the descriptor performance for scene categorization, showing good accuracy for different adjustments. Additional results show the performance of the descriptor in a navigation assistance tool.

6. REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[2] Varsha Hedau, Derek Hoiem, and David Forsyth, "Recovering the spatial layout of cluttered rooms," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[3] Bryan Russell, Alexei A. Efros, Josef Sivic, Bill Freeman, and Andrew Zisserman, "Segmenting scenes by matching image composites," in *Advances in Neural Information Processing Systems*, 2009.

[4] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba, "Recognizing scene viewpoint using panoramic place representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2695–2702.

[5] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Efficient structured prediction with latent variables for general graphical models," in *International Conference on Machine Learning (ICML)*, 2012.

[6] J. Omedes, G. López-Nicolás, and J.J. Guerrero, *Omnidirectional Vision for Indoor Spatial Layout Recovery*, pp. 95–104, Springer, 2013.

[7] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 1, pp. 36–51, 2008.

[8] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa, "Sketch-based 3d shape retrieval," in *SIGGRAPH 2010: Talks*, 2010.

[9] Jana Kosecká and Wei Zhang, "Video compass," in *European Conference on Computer Vision (ECCV)*, 2002.

[10] Lu Wang, U. Neumann, and S. You, "Wide-baseline image matching using line signatures," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[11] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] J.J. Guerrero and C. Sagüés, "Robust line matching and estimate of homographies simultaneously," *LNCS Pattern Recognition and Image Analysis*, pp. 297–307, 2003.

[13] H. Bay, V. Ferrari, and L. Van Gool, "Wide-baseline stereo matching with line segments," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*, 2005.

[14] Zhiheng Wang, Fuchao Wu, and Zhanyi Hu, "Msld: A robust descriptor for line matching," *Pattern Recognition*, vol. 42, no. 5, pp. 941–953, 2009.

[15] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

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

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[18] J. Bermúdez, L. Puig, and J. J. Guerrero, "Hypercatadioptric line images for 3d orientation and image rectification," *Robotics and Autonomous Systems*, vol. 60, pp. 755–768, 2012.

[19] Alejandro Rituerto, Ana C. Murillo, and J. J. Guerrero, "Line image signature for scene understanding with a wearable vision system," in *International SenseCam Conference*, 2013.

[20] Alejandro Rituerto, A. C. Murillo, and J. J. Guerrero, "Semantic labeling for indoor topological mapping using a wearable catadioptric system," *Robotics and Autonomous Systems*, vol. 62, no. 5, pp. 685–695, 2014, Special Issue Semantic Perception, Mapping and Exploration.