Localization in indoor environments by querying omnidirectional visual maps using perspective images

Miguel Lourenço†, Vitor Pedro† and João P. Barreto
Institute of Systems and Robotics,
Faculty of Science and Technology,
University of Coimbra.
3030 Coimbra, Portugal
{miguel, vpedro,jpbar}@isr.uc.pt

***Abstract*— This article addresses the problem of image-based localization in indoor environments. The localization is achieved by querying a database of omnidirectional images that constitutes a detailed visual map of the building where the robot operates. Omnidirectional cameras have the advantage, when compared to standard perspectives, of capturing in a single frame the entire visual content of a room. This, not only speeds up the process of acquiring data for creating the map, but also favors scalability by significantly decreasing the size of the database. The problem is that omnidirectional images have strong non-linear distortion, which leads to poor retrieval results when the query images are standard perspectives. This paper reports for the first time thorough experiments in using perspectives to index a database of para-catadioptric images for the purpose of robot localization. We propose modifications to the SIFT algorithm that significantly improve point matching between the two types of images with positive impact in the recognition based in *visual words*. We also compare the classical *bags-of-words* against the recent framework of *visual-phrases*, showing that the latter outperforms the former.**

I. Introduction

One valuable competence for a robot is the ability to localize itself with respect to the environment for performing autonomous navigation [1] and obstacle avoidance [2]. Visual recognition has been used for localization purposes by establishing correspondences between a query image and a database of geo-referenced images constituting a topological visual map [3]. However, this approach has several difficulties: (i) The query image and the corresponding image in the database, although representing the same visual contents, can substantially differ in appearance (e.g. different lightning, substantial change in viewpoint, etc); (ii) Environments containing symmetric and/or repetitive structures, e.g. doors, walls or corridors, suffer from substantial perceptual aliasing [4]; and (iii) Building a database of large scale environments can be troublesome, specially if we want an exhaustive visual coverage of the environment [3].

Omnidirectional images became widespread in the last years and are often used in many robotics applications, including vSLAM [5], visual servoing [6], and surveillance systems [7]. Panoramic cameras enable a more thorough visual coverage of the environments when compared to traditional imaging modalities due the wider field-of-view.

Indoor localization based on distinguishable scene landmarks is closely related to image retrieval [8], object recognition [9], and location recognition [3] problems. A commonly adopted scheme extracts local image features [8], quantizes their descriptors to visual words, and applies methods adapted from text search engines to accomplish visual recognition [9], [10]. Many authors take advantage of these techniques, primarily designed for perspective images, for performing image-based localization using omnidirectional images [11]. Typically the image description is accomplished by the extraction of local [8] or global [11] features for topological and metric localization using omnidirectional images in a hierarchical recognition framework [9]. In these prior works, the recognition concerns images acquired using the same type of imaging system, i.e. perspective [9], [10], or omnidirectional cameras [11].

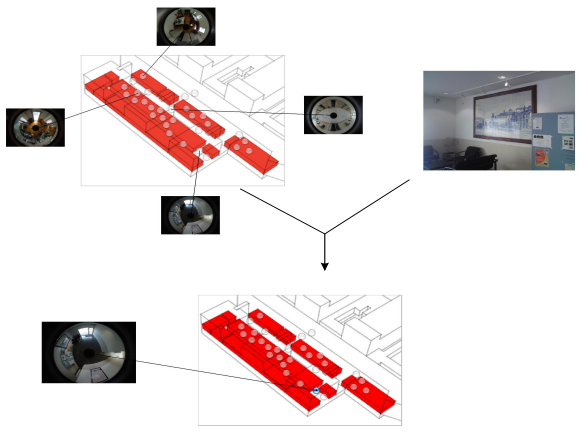


Fig. 1. Indoor localization scheme using omnidirectional visual maps.

In this paper our goal is to perform image-based localization when the query and database images are acquired using different imaging systems (hybrid imaging systems). Taking advantage of the omnidirectional images to perform

† The authors assert equal contribution and joint first authorship.

The authors acknowledge the Portuguese Science Foundation that generously funded this work through grant PTDC/EIA-CCO/109120/2008 and SFRH/BD/63118/2009. The authors was also acknowledge the Luso-American Development Foudantion that generously funded the presentation of this work at the conference.

a complete coverage of the environment, we want to retrieve the location of a query image taken from a conventional camera, e.g. a mobile robot equipped with a perspective camera, or a cell-phone image taken from a person who wants to retrieve its location. While the omnidirectional images permit to speedup the acquisition of thorough visual maps, they also introduce non-linear image distortion that increases the appearance difference between the images.

A closely related work to ours is the one of Chen *et al.* [12], where the authors perform the coverage of a city-scale outdoor environment using a panoramic camera. The authors discuss that performing matching between a perspective query and a database of omnidirectional panoramas leads to poor performance, and propose a rectification process to solve this problem. Instead of using signal reconstruction techniques, which are often subject to interpolation artifacts, we solve the problem by accounting with the distortion during keypoint detection and description. For retrieving the location of the query images we compare two approaches: the classic *bags-of-words* and the recent concept of *visual phrases* [13]. The main difference is that the *visual phrases* introduce weak spatial constraints during the recognition process, while in the standard *bags-of-words* framework the spatial layout of the features is lost.

The article outline is as follows: section II briefly reviews the SIFT algorithm, the paracatadioptric image formation process [14], [15], and strategies for matching in hybrid imaging systems [16], [17]; Section III proposes a new framework for feature detection and matching between perspectives and paracatadioptric images, and compares its performance against commonly used strategies for matching in hybrid imaging systems; section IV evaluates the proposed method in image-based indoor localization with a database of more than 100 images indexed by 450 perspective queries images. Finally, in section V, we draw conclusions and discuss future research directions.

Notation: Convolution kernels and matrices are represented by symbols in sans serif font, e.g. G, and image signals are denoted by symbols in typewriter font , e.g. $\mathtt{I}$. Vectors and vector functions are typically represented by bold symbols, and scalars are indicated by plain letters, e.g $\mathbf{x} = (x, y)^\mathsf{T}$ and $\mathbf{f}(\mathbf{x}) = (f_x(\mathbf{x}), f_y(\mathbf{x}))^\mathsf{T}$. We will use $\mathsf{U}_{(i,j)}$ to denote the entry of the i^{th} row and the j^{th} column of a matrix.

II. Background

In this section we briefly review the SIFT algorithm and the image formation model assumed along this article. We also explain how a cylindrical panorama can be built using a paracatadioptric image and discuss some strategies for matching in hybrid imaging systems.

A. Scale Invariant Feature Transform (SIFT)

Visual image location often relies on distinguishable scene landmarks (image keypoints) that can be reliable matched across views [3]. In this paper we adopt the SIFT features [8] due to its robustness to scale, rotation and small viewpoint changes [18], [19]. The keypoint detection uses a scale-space representation of the image where the Laplacian-of-Gaussian ($\mathtt{LoG}$) is approximated by Difference-of-Gaussian ($\mathtt{DoG}$) [8]. Let $\mathtt{I}(x, y)$ and $\mathsf{G}_\sigma(x, y)$ be respectively an image signal and 2D Gaussian function with standard deviation σ. The blurred version of $\mathtt{I}(x, y)$ is obtained by its convolution with the Gaussian kernel

$$\mathtt{L}_\sigma(x, y) = \mathtt{I}(x, y) * \mathsf{G}_\sigma(x, y)\,, \tag{1}$$

and the $\mathtt{DoG}$ pyramid is computed as the difference of consecutive filtered images with the standard deviation differing by a constant multiplicative factor:

$$\mathtt{DoG}_{k^{n+1}\sigma}(x, y) = \mathtt{L}_{k^{n+1}\sigma}(x, y) - \mathtt{L}_{k^n\sigma}(x, y)\,. \tag{2}$$

Each pixel in the $\mathtt{DoG}$ pyramid is compared with its neighbors in order to find local extrema in scale and space dimensions. These extrema are subsequently filtered and refined to obtain keypoints. The next step concerns the computation of the descriptor vectors using the image gradients of a local patch around each detected keypoint. Scale invariance is achieved by performing all the computations at the scale of selection in the Gaussian pyramid. The method starts by finding the dominant orientation of the local gradients, and uses it for rotating the image patch towards a normalized position. Finally, the SIFT descriptor is computed by performing a Gaussian weighting of gradient contributions, quantizing the orientations, and building histograms that accumulate magnitudes. For further details see [8].

B. Image Formation Model

Barreto and Araujo [14], [15] show that the mapping between points in the 3D world and points in the paracatadioptric image plane can be divided in three steps:

1) Visible points in the scene $\mathbf{X_h}$ are mapped into projective rays/points $\hat{\mathbf{x}}$ in the catadioptric system reference frame that is centered in the effective view point. The transformation is linear and can be described by a 3 x 4 matrix P such that
$$\hat{\mathbf{x}} = \mathsf{P}\mathbf{X_h} = \mathsf{R_c}\,[\,\mathsf{I}\,|\,-\,\mathsf{C}\,]\mathbf{X_h} \tag{3}$$
where C represents the world origin coordinates in the catadioptric system reference frame, $\mathsf{R_c}$ is the rotation matrix between the two coordinate systems, and I is a 3 x 3 identity matrix.
2) A non-linear function $\mathbf{h}$ maps points $\hat{\mathbf{x}}$ into points $\bar{\mathbf{x}}$ in a second oriented projective plane.
$$\bar{\mathbf{x}} = \mathbf{h}(\hat{\mathbf{x}}) = \begin{pmatrix}\hat{x} & \hat{y} & \hat{z} + \sqrt{\hat{x}^2 + \hat{y}^2 + \hat{z}^2}\end{pmatrix}^\mathsf{T} \tag{4}$$
3) Projective points $\mathbf{x}$ in the catadioptric image plane are obtained after the projective transformation
$$\mathbf{x} = \underbrace{\mathsf{K_c}\begin{bmatrix} 2p & 0 & 0 \\ 0 & 2p & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathsf{H_c}}\bar{\mathbf{x}} \tag{5}$$
where $\mathsf{H_c}$ depends on the mirror parameters (lactus rectum of the parabolic mirror p) and camera intrinsic parameters $\mathsf{K_c}$

C. Cylindrical Coordinates

It is possible to obtain virtual perspectives by back-projecting the omnidirectional images into planes. However, we aim at using panoramic images for recognition purposes, making use of the thorough coverage of the environment captured by a single image. For further considerations on how to obtain virtual camera perspectives we point the readers to [15]. It is also possible to map the original image into a cylinder and unfold it to obtain a panorama. Let $\hat{\mathbf{x}}$ be the backprojection of the image point $\mathbf{x}$:

$$\hat{\mathbf{x}} = (\hat{x}, \hat{y}, \hat{z})^T = \mathbf{h}^{-1}(\mathsf{H}_\mathsf{c}^{-1}\,\mathbf{x}) \tag{6}$$

The representation of $\hat{\mathbf{x}}$ in cylindrical coordinates is:

$$\begin{cases} \theta = s \cdot \arctan\left(\dfrac{\hat{x}}{\hat{y}}\right) \\ h = s \cdot \dfrac{\hat{z}}{\sqrt{\hat{x}^2 + \hat{y}^2}} \end{cases} \tag{7}$$

with s being a scaling factor (the *radius* of the cylinder). We consider $s = f$, where f is the focal length, in order to minimize the deformation near the center of the image [20]. Figure 2(b) is the result of rectifying the paracatadioptric image of Fig. 2(a). The transformation of the catadioptric image to the cylindrical panorama requires the calibration matrix H_c that we obtain using the *CatPack* toolbox made available by Barreto [21].

(a) Original image (b) Cylinder

Fig. 2. Cylindrical panorama obtained from the warping of paracatadioptric image of Figure 2(a).

D. Matching in Hybrid Imaging Systems

One possible approach to obtain matches between images coming from central catadioptric systems and conventional cameras was proposed by Luis Puig *et al.* [16]. The omnidirectional images are warped using a transformation to polar coordinates using (8). SIFT features are computed on the warped and perspective images for establishing putative matches.

$$\theta = \arctan\left(\frac{y}{x}\right) \qquad \rho = \sqrt{x^2 + y^2} \tag{8}$$

The generated polar images are very similar to the ones obtained using the mapping to cylindrical coordinates of section II-C. However, the transformation from cartesian to polar coordinates has the advantage of not requiring camera calibration.

Recently, Hansen et al. [17] proposed an extension of the SIFT algorithm for wide angle images. The method assumes that camera calibration is known and they suggest to back-project the image onto an unitary sphere and build a scale-space representation that is the solution of the diffusion equation over the sphere. Such representation minors the problems inherent to planar perspective projection, enabling RD invariance and extra invariance to rotation. However, the approach requires perfect camera calibration for both perspective and catadioptric images. In this work we assume that the perspective camera is not calibrated such that the query images can be acquired by a hand-held device, e.g. cell-phone camera.

III. Feature Extraction and Matching in Hybrid Imaging Systems

In this section we propose a new method for extracting image features from omnidirectional images that can be reliable matched with perspective image features. Instead of rectifying the omnidirectional image to perspective images [22], we implicitly compensate the distortion effect based on the rectification to cylindrical coordinates, which enables the use of the wide field-of-view of the omnidirectional images. Finally, we evaluate the proposed method using standard repeatability and precision-recall tests, and compare it against some approaches for matching between mixtures of perspectives and paracatadioptrics images.

A. Implicit cylindrical rectification - cylSIFT

1) Keypoint Detection: The objective here is to generate a scale-space representation equivalent to the one that would be obtained by filtering the cylindrical panorama. Instead of explicitly computing a new image using signal reconstruction techniques, which are often subject to interpolation artifacts [22], we adapt the convolution kernels to directly process the paracatadioptric image samples.

Through the manipulation of (6) and (7), we can re-write the mapping from paracatadioptric coordinates to cylindrical coordinates as

$$\mathbf{u} = \mathbf{f}^{-1}(\mathbf{x}) = \begin{pmatrix} f_u^{-1}(x,y) \\ f_v^{-1}(x,y) \end{pmatrix} = \begin{pmatrix} f \cdot \arctan(x/y) \\ \dfrac{f^2 - r^2}{2r} \end{pmatrix}. \tag{9}$$

The inverse of (9) provides the mapping between cylindrical and paracatadioptric coordinates:

$$\mathbf{x} = \mathbf{f}(\mathbf{u}) = \begin{pmatrix} f_x(u,v) \\ f_y(u,v) \end{pmatrix} = \begin{pmatrix} y\tan(\dfrac{u}{f}) \\ \cos\left(\frac{u}{f}\right)\left(\sqrt{f^2+v^2} - v\right) \end{pmatrix}. \tag{10}$$

Consider the convolution of the cylindrical image $\mathtt{I}^{cyl}$ with a Gaussian kernel with standard deviation σ. By writing the convolution operation of (1) explicitly, it comes that the blurred image is

$$\mathtt{L}_\sigma^{cyl}(s,t) = \sum_u \sum_v \mathtt{I}^{cyl}(u,v)\,\mathsf{G}_\sigma(s-u, t-v)\,. \tag{11}$$

If $\mathtt{I}$ is the paracatadioptric image, then from the mapping relation (10) it follows that $\mathtt{I}^{cyl}(\mathbf{u}) = \mathtt{I}(\mathbf{x})$, with $\mathbf{x} = \mathbf{f}(\mathbf{u})$. Replacing $\mathtt{I}^{cyl}$ by $\mathtt{I}$ and switching the variables (u, v) by (x, y) using (9), we obtain the result of (12). This equation

$$\mathtt{L}_{\sigma}^{cyl}(s,t) \;=\; \sum_{x}\sum_{y} \mathtt{I}(x,y)\, \mathsf{G}_{\sigma}(s - f_u^{-1}(x,y), t - f_v^{-1}(x,y)) \tag{12}$$

$$\mathtt{L}_{\sigma}(h,k) \;=\; \sum_{x}\sum_{y} \mathtt{I}(x,y)\, \mathsf{G}_{\sigma}\Big(f \cdot \Big(\arctan\Big(\frac{h}{k}\Big) - \arctan\Big(\frac{x}{y}\Big)\Big),\; \frac{f^2(\delta-1) + \delta r^2(\delta-1)}{2\delta r}\Big) \tag{13}$$

computes the blurred image $\mathtt{L}_{\sigma}^{cyl}$ directly from the original distorted frame $\mathtt{I}$.

Let's now apply distortion to the blurred image $\mathtt{L}_{\sigma}^{cyl}$ in order to obtain $\mathtt{L}_{\sigma}$. This can be achieved in an implicit manner using the previous mapping functions. After replacing the cylindrical image coordinates (s,t) by their paracatadioptric counterpart (h,k) and performing some algebraic simplifications, we obtain the adaptive filtering of (13) with r being the distance between the center and the image location where the filter is applied

$$r \;=\; \sqrt{h^2+k^2}\,, \tag{14}$$

and δ being the ratio between the radius d of each pixel contribution and r

$$\delta \;=\; \frac{d}{r} \;=\; \frac{\sqrt{x^2+y^2}}{\sqrt{h^2+k^2}}\,.$$

Note that now the smoothing convolution is an operation of $\mathbb{R}^2 \times \mathbb{R}^4 \rightarrow \mathbb{R}_+$ due to its dependence in (h,k) and (x,y). For each radius, the adaptive blurring kernel has the same shape, but with different orientations (see Fig.3(a)).

It is well known that the standard Gaussian filter is a rank 1 matrix that can be written as the outer product of two 1D gaussian filters of the same standard deviation. This permits to implement the convolution process separately in X- and Y-directions, which permits to considerably speedup the smoothing process [23]. Instead of computing the *cylindrical* Gaussian for each image pixel position, we approximate (13) by the closest rank 1 Gaussian filter estimated using Singular Value Decomposition

$$\begin{bmatrix}\mathsf{U} & \mathsf{S} & \mathsf{V}\end{bmatrix} = \mathtt{SVD}(\mathsf{G}_{\sigma}). \tag{15}$$

Thus, the rank 1 Gaussian kernel that better approximates G_{σ} is

$$\mathsf{G}_{\sigma,rank=1} = \mathsf{U}_{(:,1)}\mathsf{S}_{(1,1)}\mathsf{V}_{(:,1)}{}^{\mathsf{T}} \tag{16}$$

with the filter being accordingly normalized to have unit sum (see Fig.3(a)). We have observed that this decomposition has two significant advantages: (i) For every image radius the same $\mathsf{G}_{\sigma,rank=1}$ can be used, which enables separable convolution for each radius in a similar way to [22]; and (ii) a filter bank can be computed offline and loaded into memory when required. We consider the same filter bank for all the paracatadioptric images used throughout this paper.

2) Keypoint Description: Concerning the SIFT descriptor computation we can explicitly correct the image gradients, by warping the image to the cylinder and computing the gradients in this reconstructed signal, or implicitly correct them by measuring the gradients in the original image and correct the result using the derivative chain rule. The implicit approach avoids the propagation of interpolation artifacts inherent to the image re-sampling [22].

Let $\mathtt{I}$ be the catadioptric image and $\mathtt{I}^{cyl}$ be the cylindrical panorama. The mapping relation between the two images is the following:

$$\mathtt{I}^{cyl}(\mathbf{u}) \;=\; \mathtt{I}(\mathbf{f}(\mathbf{u}))\,.$$

Applying the derivative chain rule it yields

$$\nabla \mathtt{I}^{cyl} \;=\; \mathsf{J}_{\mathbf{f}}\,.\,\nabla \mathtt{I} \tag{17}$$

with $\nabla \mathtt{I}^{cyl}$ and $\nabla \mathtt{I}$ being respectively the gradient vectors in $\mathtt{I}^{cyl}$ and $\mathtt{I}$, and $\mathsf{J}_{\mathbf{f}}$ being the 2×2 Jacobian matrix of the mapping relation given in (10). The Jacobian matrix can be written in terms of paracatadioptric image coordinate $\mathbf{x} = (x,y)^{\mathsf{T}}$:

$$\mathsf{J}_{\mathbf{f}} \;=\; \begin{pmatrix} \frac{r^2}{fy} & 0 \\ -\frac{x}{fr}\Big(\tau + \sqrt{\tau^2+f^2}\Big) & \frac{y}{r}\Big(\frac{\tau+\sqrt{f^2+\tau^2}}{\sqrt{f^2+\tau^2}}\Big) \end{pmatrix}$$

with r denoting the radius of $\mathbf{x}$ and $\tau = \frac{r^2-f^2}{2r}$.

In summary, we propose to measure the gradients directly in the original distorted image $\mathtt{I}$, evaluate the jacobian matrix $\mathsf{J}_{\mathbf{f}}$ at every relevant pixel location, and correct the gradient vectors $\nabla \mathtt{I}$ using the differential chain rule of (17). The final descriptor is generated from the undistorted gradients $\nabla \mathtt{I}^{cyl}$ following the procedure described in section II.

B. Performance evaluation

1) Methods under evaluation: In this hybrid matching comparison, SIFT [8] is always used to extract features in the perspective images and the test only differ in terms of the method used to extract features in the paracatadioptric/rectified views. We compare the proposed cylSIFT method against the following approaches: Application of SIFT over (i) paracatadioptric images (SIFT); (ii) rectification to polar coordinates (*Polar*); (iii) rectification to cylindrical coordinates (*Cylinder*); and (iv) virtual image perspectives (*VCP*). To generate the VCP we manually select the region in the omnidirectional images that correspond to the visual contents of the perspectives. Without this prior knowledge we would need to render 4 or more perspectives for each omnidirectional image, and still be subject to viewpoint changes arising in the synthetical generated perspective images. Although the VCP is not a direct competitor of our method because it does not capsulate the same wide field-of-view in one image, it is the theoretical top performer since matching is accomplished between images with no distortion,

(a) Adaptive filters (b) Set A (c) Set B (d) Set C (e) Set D

Fig. 3. Example of the data sets used for detection and description evaluation.

being included in the performance evaluation study for a sake of completeness of the study.

2) Metrics for evaluation: In terms of detection evaluation, the repeatability of keypoint detection is unarguably the most important property of a reliable detector [18]. Let's consider S_{cata} and S_{pers} as being the set of keypoints detected in the paracatadioptric image (or rectifications obtained from it) and perspective images, respectively. Given two images of the same scene, the repeatability measures the percentage of the features detected on the scene part visible in both images:

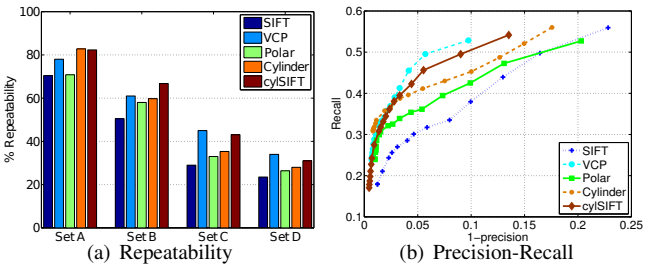
$$\%_{\text{Repeatability}} = \frac{\#(S_{cata} \cap S_{pers})}{\#S_{pers}} * 100 \qquad (18)$$

where $\#$ denote the cardinality of the sets. For matching evaluation we use the traditional 1-precision *vs* recall curves [19]. The recall indicates the percentage of correct matches M^{true} obtained over the entire set of possible correct matches $S^{true} = S_{cata} \cap S_{pers}$. This metric must be complemented by the precision that measures the percentage of correct matches over the entire set of matches M, i.e. the precision measures how well the algorithm discards keypoints that have no correspondence. The curves are obtained by varying the threshold λ of the similarity distance between the descriptors [8], [19].

$$\text{recall}(\lambda) = \frac{\#M^{true}}{\#S^{true}} \quad \text{precision}(\lambda) = \frac{\#M^{true}}{\#M} \qquad (19)$$

3) Datasets: We collected 13 paracatadioptric images taken in different places using a camera with a resolution of 2272×1704. On the perspective side we collect 4 different perspective image sets (Fig. 3): set (A) was acquired fronto-parallel to the scene, at the same location of the paracatadioptric system; set (B) was acquired from the same position as the paracatadioptric image and with an angle of approximately 45 degrees between the optical axis and the vertical plane; set (C) presents strong scale changes while preserving the fronto-parallel viewpoint; and set (D) was taken from different positions and viewpoints relatively to the paracatadioptric images, to test strong viewpoint changes. The resolution of the perspective images is 1600x1200. It is important to notice that, at this stage, we only consider images of planar scenes that enables to find a ground truth homography[1] for verification of detection and matching results.

[1]The ground truth homography is computed after rectifying the paracatadioptric coordinates to perspective coordinates.



(a) Repeatability (b) Precision-Recall

	Panoramic Image				Persp.
	SIFT	Polar	Cylinder	cylSIFT	VCP
Detections	1328	1482	1613	1433	401
Time (sec)	4.4	7.78	8.05	5.6	2.4^{+}
No . Matches	94.6	96.2	112.1	120.8	126.9

Fig. 4. Detection and description evaluation in planar image pairs. Fig. 4(a) compares the repeatability scores of the several methods evaluated, while Fig. 4(b) concerns description evaluation. We can observe that using the cylSIFT approach permits to have similar scores to the rectification for a perspective view. Additionally, we provide the average running time of every method, number of detection and number of matches established using the similarity distance thresholded at 0.9. The computation differences between the SIFT and the cylSIFT rely on the offline computation of the filter bank, which in our Matlab implementation takes in average 1 second, and in the gradient correction technique. In VCP, Polar and Cylinder the rectification process using our Matlab routines is included. $(\cdot)^{+}$ denotes that for the VCP we only show the running the time for the correct perspective. In practice at least 4 perspective images must be rendered for each omnidirectional to cover its wide field of view.

4) Results and discussion: The repeatability of detection and precision-recall curves for description can be observed in Fig. 4. We can observe that the cylSIFT performs better than most competing methods over the panoramic images. The image re-sampling for distortion compensation requires the reconstruction of the discrete image signal. This reconstruction process can either remove high frequency components and/or introduce new spurious frequencies [23], being highly prejudicial in the detection step [22]. The Polar and the Cylinder generate similar images and it is expected that both provide similar results. However, as the rectification to cylindrical coordinates uses the calibration matrix and the non-linear function characteristic of the mirror, the mapping of the latter is more accurate than the former, which explains the observed better performance. The cylSIFT method performs very closely to the VCP approach, showing that, even dealing with the distortion on the cylinder, the cylSIFT is capable of performing close the perspectives generated through interpolation.

In terms of description, we can observe that performing implicit gradient correction provides gains in terms of matching performance, when compared with the other descriptors

computed in the panoramic images. Once more it is verified that the VCP provides the best matching scores, which is expected since the description space, although subject to interpolation artifacts, does not present any non-linear distortion.

In summary, we can conclude that the cylSIFT outperforms SIFT applied directly over the omnidirectional image, as well as polar and cylindrical panoramas. The VCP approach outperforms the cylSIFT algorithm due to the correct alignment between the perspective image and generated VCP (see results of set D in Fig. 4(a)). In a real application scenario, this correct alignment is not known in advance, precluding a good performance for this method.

IV. Image-based Localization using Hybrid Imaging Systems

In this section we evaluate the proposed cylSIFT method for image-based localization. Given a query image, acquired with a standard camera (e.g. robot or a person with a conventional camera), the localization is obtained by searching and retrieving the most similar view in a database of omnidirectional visual maps.

A. Retrieval schemes

In our retrieval application, we compare two different searching approaches. The first method uses the standard *bags-of-visual words* (*BoV*) approach. A vocabulary tree is built using hierarchical *k*-means clustering, with *k* defining the branch factor of the tree. Each branch is recursively split into *k* new groups along *L*-levels of the tree, which totalizes k^L visual words. The correspondence between images is given by measuring the similarity between the visual words in a query image and in the database images [9]. Although this scheme provides good performance in several recognition scenarios [9], [11], it discards the spatial relation of the visual words during retrieval that can be relevant to disambiguate situations of perceptual aliasing [3]. The second method uses the new concept of *visual phrases* (*GVP*) [13]. The objective of using GVP is to take into account the spatial relations between visual words. For each pair of the same word in the query and database images, the offset is computed by subtracting their corresponding locations. A set of *n* visual words in a certain spatial layout define a GVP of length *n*. The image space is quantized into cells to tolerate shape deformation and to build an efficient voting scheme. After computing the offset, a vote is generated on the offset space. *n* votes in the same offset cell correspond to a co-occurring GVP of length *n*. For further details see [9], [13].

B. Feature Extraction Methods and Database considerations

The extraction of features in the query images is always performed using the standard SIFT algorithm. On the database side, we consider the following features extraction and description schemes: SIFT applied over (i) the paracatadioptric images; (ii) the cylindrical rectification *Cylinder* and (iii) virtual image perspectives *VCP*; and (iv) the cylSIFT features computed in the paracatadioptric images.

The feature extraction techniques and searching schemes are tested by performing queries on a database of 118 paracatadioptric images that provide a detailed visual map of the building where the robot operates. Concerning the VCP database, we render 4 perspectives for each omnidirectional image. Each generated perspective image has a field of view of 108° and resolution of 1600×1200. Unlike in the tests of section III, the VCP images are generated in an unsupervised manner, meaning that each omnidirectional image gives raise to 4 VCP without assurance that one of the VCP is aligned with the perspective query image. We use 451 query images for evaluating which combination of retrieval scheme, vocabulary size and feature extraction method (at the database side) performs better for the task.

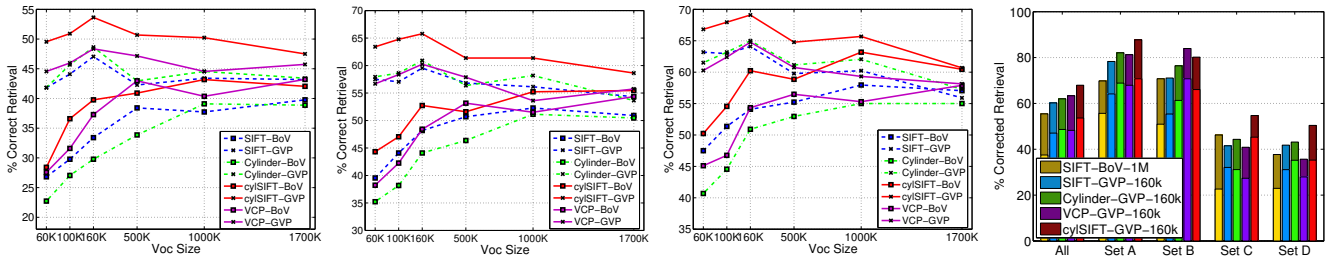
The performance of retrieval is given by the percentage of correctly retrieved locations in first place (Top 1), and in the sets of 3 and 5 images with highest scores (Top 3 and Top 5). Finally, the best retrieval method for each feature extraction technique is selected and the top 5 images are re-ranked through geometrical verification within a RANSAC framework [24].

C. Results and Discussion

Figure 5 presents the retrieval results. The cylSIFT approach is the one providing the highest retrieval scores, independently of the searching scheme and vocabulary size. Increasing the number of words in the vocabulary increases the performance of the BoV approach. In this case, the recognition is performed in a word-by-word basis and more discriminative words tend to provide better retrieval results. We also observed that using a lower vocabulary size tends to favor the performance of the GVP. Since vocabularies of small sizes are less discriminative more common words between the query and the database image can be established and more visual phrases exist.

For each feature extraction method, we selected the best retrieval scheme (GVP with a vocabulary size of $160k$) and performed re-ranking on the top 5 images using strong geometric constraints within a RANSAC framework (Fig. 5(d)). In addition to the average correct retrieval scores, we also provide the results for each perspective sets. The higher quality of matching provided by our method can be clearly seen for all the 4 sets, but with particular emphasis in the most difficult ones (set C and D). It is also important to notice that in set D the VCP tends to be outperformed by all other methods. This is due to the fact that in general the perspective image is misaligned with the generated VCP, meaning that such schemes is only effective if we known in advance which region of the omnidirectional image is being viewed in order to ensure small changes in viewpoint.

One important observation is that the interpolation used in the explicit cylindrical and VCP images has a negative impact in the visual words discriminability. Using the BoV with descriptors extracted in cylindrical images does not lead to an increase in performance when comparing with SIFT, showing that the visual words computed on these descriptors are less discriminative. While in the section III, two descrip-



(a) Top 1 (b) Top 3 (c) Top 5 (d) RANSAC re-ranking

Fig. 5. Retrieval results. We have tested several combinations regarding feature extraction techniques, vocabulary size and searching scheme. The cylSIFT method proved to be superior to the other feature extraction approaches, regardless of the type of retrieval scheme and vocabulary size. For each feature extraction method, we selected the best retrieval scheme and performed re-ranking on the top 5 images using strong geometric constraints within a RANSAC framework (Fig. 5(d)).The darker colors represent the improvement obtained over the GVP framework, with a vocabulary size of 160k. We additionally include the scores of a naive approach to the problem where SIFT features and standard BOV are used for localization recognition.

tors were considered a match by using similarity distance (nearest neighbor distance ratio) [19], in the vocabulary tree, two descriptors belong to the same visual word if they are close to the same centroid. Therefore, the smaller the euclidean distance between two descriptors, the greater the probability of belonging to the same visual word. Although the interpolation artifacts do not have a large influence in the nearest neighbor ratio, they seem to be particularly relevant for the computation of the image visual words. The implicit filtering approach seems to be immune to this phenomena and takes full advantage of its higher matching performance.

V. Conclusions

This paper focus on indoor image-based localization by querying omnidirectional maps using perspectives. We take advantage of the wide field-of-view of the images, which enable a complete description of the environment with minimum effort. To successfully retrieve the omnidirectional image using a perspective, we develop a new algorithm for feature detection and description based on the rectification to cylindrical images. Extensive experiments prove that our method outperforms explicit image rectification methods, proving to be beneficial for image-based localization by improving the rate success rate in $10-15\%$. We also compare the classic *bags-of-words* against the recent *visual phrases*, showing that the latter significant improves the recognition scores. In the future we will extend our omnidirectional visual map, and make it available for the community to stimulate further research on the topic.

References

[1] S. Se, D. Lowe, and J. Little, "Vision-based Mobile Robot Localization And Mapping using Scale-Invariant Features," in *IEEE Int. Conf. on Robot. and Autom.*, 2001.

[2] A. Chavez and D. Gustafson, "Vision-Based Obstacle Avoidance Using SIFT Features," in *Int. Symp. on Adv. in Visual Comput.*, 2009.

[3] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *Int. J. Robot. Res.*, 2008.

[4] F. Werner, F. D. Maire, J. Sitte, H. Choset, S. Tully, and G. Kantor, "Topological slam using neighbourhood information of places," in *IEEE-IROS, Int. Conf. on Intell. Robot. and Syst.*, 2009.

[5] D. Scaramuzza, F. Fraundorfer, M. Pollefeys, and R. Siegwart, "Closing the Loop in Appearance-Guided Structure-from-Motion for Omnidirectional Cameras," in *Work. on Omni. Vis., Cam. Networks and Non-class. Cameras* , 2008.

[6] G. L. Mariottini and D. Prattichizzo, "Image-based visual servoing with central catadioptric cameras," *Int. J. of Robot. Research*, vol. 27, 2008.

[7] A. Pretto, E. Menegatti, and E. Pagello, "Omnidirectional dense large-scale mapping and navigation based on meaningful triangulation," in *IEEE Int. Conf. on Robot. and Autom.*, 2011.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, 2004.

[9] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Int Conf. on Comput. Vis. & Patt. Recog."*, 2006.

[10] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE Int Conf. on Comput. Vis.*, 2003.

[11] J. K. A. C. Murillo, P. Campos and J. J. Guerrero, "GIST vocabularies in omnidirectional images for appearance based mapping and localization," in *Work. on Omni. Vis., Cam. Networks and Non-class. Cameras* , 2010.

[12] D. Chen, G. Baatz, K. Koeser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *IEEE Int Conf. on Comput. Vis. & Patt. Recog."*, 2011.

[13] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *IEEE Int Conf. on Comput. Vis. & Patt. Recog."*, 2011.

[14] J. Barreto and H. Araujo, "Issues on the geometry of central catadioptric image formation," in *IEEE Int Conf. on Comput. Vis. & Patt. Recog.*, 2001.

[15] J. P. Barreto, "A Unifying Geometric Representation for Central Projection Systems," *Comput. Vis. Image Underst.*, vol. 103, 2006.

[16] J. J. G. L. Puig and P. Sturm, "Matching of omnidirectional and perspective images using the hybrid fundamental matrix," in *Work. on Omni. Vis., Cam. Networks and Non-class. Cameras* , 2008.

[17] P. Hansen, P. Corke, and W. Boles, "Wide-Angle Visual Feature Matching for Outdoor Localization," *Int. J. of Robot. Research*, vol. 29, 2010.

[18] 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. Vision*, vol. 65, 2005.

[19] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. On Patt. Analysis and Mach. Intell.*, vol. 27, October 2005.

[20] R. Szeliski, "Image alignment and stitching: a tutorial," *Found. Trends. Comput. Graph. Vis.*, vol. 2, 2006.

[21] J. P. Barreto and H. Araujo, "Geometric properties of central catadioptric line images and their application in calibration," *IEEE Trans. On Patt. Analysis and Mach. Intell.*, vol. 27, 2005.

[22] M. Lourenco, J. Barreto, and F. Vasconcelos, "sRD-SIFT Keypoint Detection and Matching in Images with Radial Distortion," *IEEE Trans. on Robotics.*, 2011.

[23] L. Velho, A. Frery, and J. Gomes, *Image Processing for Computer Graphics and Vision*. Springer London, 2008.

[24] M. A. Fischler and R. C. Bolles, "RANdom SAmple Consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, 1981.