FCTUC

University of Coimbra
Faculty of Sciences and Technology
Department of Electrical and Computer Engineering

Visual recognition for localization purposes using omnidirectional images maps

Vítor Manuel Castela Pedro

Coimbra, 2011

Visual recognition for localization purposes using omnidirectional images maps

Advisor: Prof. João Pedro Barreto

Committee:

Prof. Dr. Jorge Dias

Prof. Dr. Gabriel Falcão

Prof. Dr. Joao P. Barreto

A Thesis submitted for obtaining the degree of Integrated Master
in Electrical and Computer Engineering

Department of Electrical and Computer Engineering

Faculty of Sciences and Technology,
University of Coimbra

September 2011

Acknowledgements

During this semester many people contributed to this work.

In first place i want to thank my parents and my brother for the unconditional support all these years and for always believing in me.

Thanks to my advisor Prof. João Pedro Barreto, for all the good advices and for keeping me in the right way. Thanks for the great research environment and for always being available.

For my lab colleagues: Miguel, Michel, Melo, Pinto, África, João and Aniana a big thank you for all the good moments and support. I want to specially thank my colleague and friend Miguel, without him, i could not have achieved my work with success. Thanks for always being there to help.

Thanks to João Nuno, Heitor, Zélia, Cátia Vanessa, Liliana, Nuno Lucas, Ricardo Lopes, Pascoal and Palhinhas for all the support and true friendship. *Obrigado João Pascoal por acreditares no meu trabalho aqui em Coimbra :).* Thanks to Aniela, for the polish meal!

The last but not the least, i want to acknowledge my girlfriend Sílvia, for all the support and for standing by my side when i needed the most.

Thanks to all!

Contents

List of Figures

List of Tables

Abstract

This thesis is an exploratory work with the objective of developing techniques for recognition using visual maps constituted by paracatadioptric images. This kind of images contain a more complete description of the scene, covering a wider area than e.g. perspective images. However, the radial distortion they present is high which difficult their usage. The main challenge of the thesis is establishing correspondences between a query image captured by a standard camera and the paracatadioptric images.

Different types of rectification strategies are studied in order to correct the radial distortion present in the paracatadioptric images. Matching between perspective images and rectified images from the paracatadioptric images is performed, using SIFT algorithm [18]. Also, a simple procedure to roughly estimate the calibration matrix of the system is explained. Additionally, several modifications to the original SIFT algorithm are proposed, including a change in the initial scale for the Gaussian blurring and a novel approach for feature detection and description of stable local features, directly over the paracatadioptric images (based on [17]). With this approach, the detection is carried in a scale-space image representation built using an adaptive Gaussian filter that takes into account the geometry of the paracatadioptric system. As an additional result, a brief study of feature detection and matching between paracatadioptric images is performed. A new mapping function for the adaptive filtering is tested, outperforming the other approaches.

Another important topic of the thesis is to perform image based localization using a database of omnidirectional images. A recognition scheme using an hierarchical vocabulary tree is built, based on [25]. Different visual vocabularies and training data are used. Additionally, several methods for feature extraction in the database images are analyzed, including the original SIFT algorithm and SIFT with implicit filtering. Finally, methods to improve the retrieval performance are studied. To encode more spatial information in the searching step, the concept of geometry-preserving visual phrases is used [33]. Additionally, to provide a more precise ranking of the retrieved images, a geometric consistency check (using RANSAC) is performed on the top-ranked images.

Chapter 1

Introduction

Localization plays an important role in a wide range of robotic applications, e.g. autonomous navigation and obstacle avoidance, and also in people's life. A possible approach to achieve localization is to use visual recognition. Localization based on visual information has benefits in terms of cost and flexibility, e.g. localization systems like GPS have limitations when used indoors. The principal component of image based localization is the search for the most similar view in an image database representing the environment. Visual recognition can be used for localization purposes by establishing correspondences between a query image and a database of geo-referenced images constituting a topological visual map. This approach has, however, several difficulties:

(i) The query image and the corresponding image in the database, although representing the same visual contents, can differ substantially in appearance (e.g. different lightning, substantial change in viewpoint, etc).

(ii) Perceptual aliasing [13]. If the environment contains symmetric and/or repetitive structures, e.g. doors, walls or corridors, then it leads to perceptual ambiguities.

(iii) Building the database of large scale environments can be troublesome, specially if we want an exhaustive visual coverage of the environment.

Omnidirectional images became widespread in the last years and are often used in

robotics and surveillance enabling panoramic imaging. This kind of images have one major advantage when compared to traditional imaging modalities: one image contains a more complete description of the scene, since it is able to cover a wider area than e.g. perspective images. This decreases the number of images in the database, ergo, memory requirements and time to construct the database. Therefore, we propose to use omnidirectional images to overcome problem (iii). Unfortunately, this option aggravates the issue (i) due to the radial distortion present in these kind of images, which considerably difficult their usage.

The objective of the thesis is to develop techniques for localization through visual recognition using visual maps constituted by geo-referenced omnidirectional images. Given a query image, acquired with a standard camera (e.g. cellphone), the localization is obtained by searching and retrieving the most similar view in the database of images (see Figure 1.1). The retrieval issue will be addressed, either by improving keypoint matching between perspective and omnidirectional images, and working with a searching scheme based on visual words.

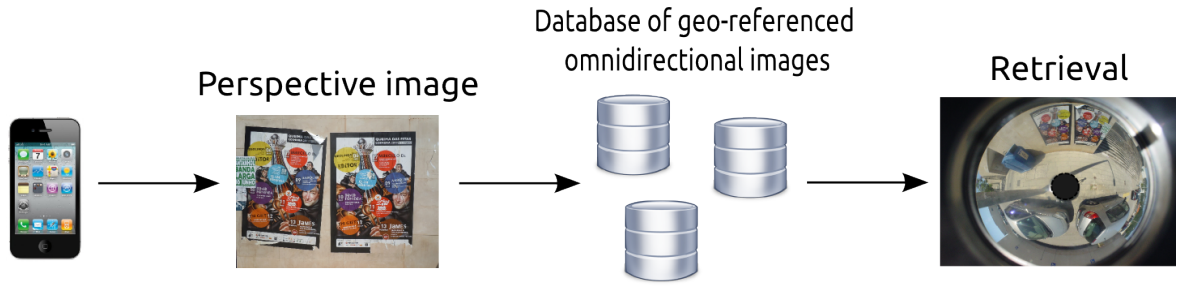


Figure 1.1: Illustration of the visual recognition scheme.

1.1 Related Work

There are previous works on localization using omnidirectional images, for instance, A. C. Murillo *et al.* [2] proposed to use SURF (Speeded-Up Robust Features) features [7] integrated with a vision-based algorithm that allows both topological and metric localization using omnidirectional images in a hierarchical approach. Also, in [3], Murillo *et al.* studied the efficiency and potential of global gist descriptor [26] adapted to catadioptric systems, and presented a new hierarchical approach for topological mapping and

localization using omnidirectional images. The gist descriptor is a global descriptor of an image representing the dominant spatial structures of the scene captured. The use of global descriptors has disadvantages like lower invariant properties and less robustness to occlusions, and advantages like more compactness in the representation of the image allowing enhancements in storage and computation efficiency. In [3], they segment the omnidirectional image in 4 symmetric parts, rotating them to a canonical orientation and computing the gist descriptor in each part. In [16], it is described a method for spatial representation, place recognition and self-localization using omnidirectional images. The spatial representation is built up invariant signatures using an adaptation of Haar invariant integrals to the particular geometry of catadioptric omnidirectional sensors. It is important to refer that all these works concern retrieval where both query and database images are omnidirectional. However, this thesis is focused in retrieval where the query is a perspective image.

When dealing with a large database of images, the problem of efficiently search for a matching image can become difficult. In the literature, very good results were shown demonstrating content based image retrieval using local scale-invariant features with various techniques of indexing and quantization (e.g. vocabulary trees) [25, 29]. In [25], the local features are hierarchically quantized in a vocabulary tree, allowing a larger and more discriminatory vocabulary to be used efficiently. The main objective is to build an indexing mechanism that enables extremely efficient retrieval. In [28], a place recognition scheme based on omnidirectional images is employed for loop detection. The scheme works with a visual word based approach. Once again, this work concerns retrieval using omnidirectional images as query and in the database.

In order to search and retrieve the most similar image in the database, correspondences between the query and the database images must be established. Image features (also called keypoints) are points of interest that can be extracted from images, and are well suited to be matched with features from other images of the same visual content. Therefore, feature matching is relevant for the retrieval problem. There is a large variety of methods and algorithms for keypoint detection and matching. SIFT algorithm, originally proposed by David Lowe [18], is one of the most robust approaches in the literature in terms of scale, rotation and minimal viewpoint invariance. Additionally, SIFT

features are highly distinctive, which is appropriate for exact matching against a large collection of features. Unfortunately, SIFT do not take into account the implicit geometry of the mirrors in catadioptric systems, penalizing the performance of the image analysis applications that directly use the omnidirectional images [10, 20] (SIFT framework was designed to be used in images that obey the standard pin-hole model). Hansen *et al.* [11] proposed an approach to extend SIFT for the case of wide angle images. They suggest back-projecting the image on an unitary sphere and building a scale-space representation that is the solution of the heat diffusion equation over the sphere, which is implemented in the frequency domain using spherical harmonics. Such transformation corrects the radial distortion and enables extra invariance to rotation. However, the approach requires perfect camera calibration and is complex and computationally expensive. In [17], Lourenço *et al.* showed that building the scale-space representation using an adaptive Gaussian filter can do better for images with radial distortion. In this thesis, we will try to extend this approach for paracatadioptric images. In [15], Luis Puig *et al.* present an hybrid matching system, mixing images coming from central catadioptric systems and conventional cameras. First, they unwarp the catadioptric images to polar coordinates in order to obtain an initial matching. Then, a robust estimation gives an estimation of the hybrid fundamental matrix and allows to detect wrong matches. Recently, Z. Arican and P. Frossard proposed a method to compute scale invariant features in omnidirectional images [4]. They developed a novel scale-invariant feature detection framework for omnidirectional images that can be mapped on the sphere. They also present a new descriptor and feature matching solution for this kind of images. Additionally, is showed that the proposed framework also permits to match features in images with different geometries.

1.2 Contributions

This thesis is an exploratory work with the objective of developing techniques for recognition using visual maps constituted by geo-referenced paracatadioptric images (topological maps). The main challenge is to develop techniques for establishing correspondences between the query image taken with a standard camera (described by the pin-hole model) and the paracatadioptric images. Therefore, different approaches for feature matching

between perspective and paracatadioptric images are studied, including a new method for detection and description of interest points that takes into account the geometry of these images. Also, a searching scheme based on a vocabulary tree structure is implemented, to efficiently search the visual map. The use of a vocabulary tree is important to improve the efficiency when handling a database containing a large number of images. This approach is complemented by using the spatial information of the features in the searching scheme, based on the concept of geomerty-preserving visual phrases.

1.3 Organization of the Thesis

This thesis is organized as follows:

- In the next Chapter, feature detection and matching, using SIFT algorithm, is performed under different conditions. Matching between perspective images and rectified images from the catadioptric images is performed (the radial distortion is corrected and the catadioptric images are approximated by a pin-hole projection). Different types of rectification strategies will be used and compared in order to select the most appropriate one.

- In Chapter 3, methods to improve the matching between standard images and catadioptric/rectified images are explained, including an algorithm for feature detection and description in paracatadioptric images. The objective is to build the scale-space representation using an adaptive Gaussian filter following [17]. The algorithm is compared against: (i) applying the original SIFT directly in catadioptric images, (ii) in rectified images after radial distortion correction, (iii) with the unwarping method proposed by Luis Puig *et al.* [15], where points in the catadioptric image are converted to polar coordinates.

  Finally, several methods for feature detection and matching between paracatadioptric images will be studied and analyzed.

- In Chapter 4, the concept of vocabulary tree is studied. The main objective is to develop a recognition scheme that scales efficiently to large databases of images. The approach of David Nistér and Henrik Stewénius [25], where local region descriptors

are hierarchically quantized in a vocabulary tree will be implemented and tested. Different visual vocabularies and training data is used and compared in order to find the most suitable searching scheme.

Additionally, an approach that can encode more spatial information in the searching step and that is efficient to be applied in large databases is studied. This approach [33] uses geometry-preserving visual phrases (GVP). A GVP is constituted by a group of visual words in a particular spatial layout. This method can provide a better initial ranking with more spatial information.

Finally, a *post-processing* step is added providing a more precise ranking of the retrieved images through a geometric consistency check.

- In chapter 5, conclusions and an outlook about this work are reported.

Chapter 2

Feature Matching in Mixtures of Perspective and Paracatadioptric Images (Hybrid)

Image features, henceforth called keypoints, are points of interest that can be identified and described in images, and are well suitable to be matched with keypoints from other images of the same scene/object captured under different acquisition conditions. Therefore, these keypoints (ideally) need to be invariant to changes in illumination, scaling, view point and rotation. Image features have a wide range of applications in Computer Vision, like object or scene recognition, stereo correspondence, and motion tracking.

Matching between images from conventional cameras (perspective images) and central catadioptric systems[1] (e.g. paracatadioptric images) is problematic because the latter have strong radial distortion. Other problems that can occur are great differences in viewpoint and rotation between the images. Additionally, most keypoints detectors and descriptors are designed for images that obey to the pin-hole model, which is not the case of catadioptric images. Therefore, and because features play a major role in recognition, one of the main objectives of this thesis is to explore techniques to improve feature matching between perspective and catadioptric images.

[1]Central catadioptric systems provide a wide field of view (FOV) while keeping an unique projection center.

2.1 SIFT features

There is a large variety of different methods and algorithms for detecting and matching keypoints across views [18, 19, 22]. Scale Invariant Features Transform (SIFT) algorithm, proposed by David Lowe [18], is a well known method for identification and description of distinctive invariant features that was designed to be stable and efficient. These features are invariant to image scaling and rotation of the image plane, and partially invariant to changes in illumination and camera viewpoint. Several studies showed that SIFT algorithm is one of the most robust techniques for keypoint detection and description [23, 24]. Due to this, SIFT is used in this thesis for the task of feature extraction and description.

2.1.1 SIFT Pipeline

The SIFT algorithm can be divided in two main steps:

- **Scale-space extrema detection:**

  The first step for feature detection is to identify keypoint locations and scales that can be repeatably extracted under different views of an object. To be able to detect keypoints invariant to scale changes and with high repeatability rates, a multi-scale approach is needed. Therefore, in order to efficiently detect stable keypoint locations in scale-space, Lowe performs extrema detection in the difference-of-Gaussian function, $\mathsf{D}(x, y, \sigma)$. The scale-space representation $\mathsf{L}(x, y, \sigma)$, with σ denoting the scale, is obtained from the convolution of a variable-scale Gaussian, $\mathsf{G}(x, y, \sigma)$, with an image, $\mathtt{I}(x, y)$:

  $$\mathsf{L}(x, y, \sigma) = \mathtt{I}(x, y) * \mathsf{G}(x, y, \sigma) \tag{2.1}$$

  where $*$ is the convolution operation in x and y, and

  $$\mathsf{G}(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.2}$$

  Difference-of-Gaussian images can be computed as:

$$\mathsf{D}(x, y, \sigma) = \mathsf{L}(x, y, k\sigma) - \mathsf{L}(x, y, \sigma) \tag{2.3}$$

where k is a multiplicative constant. The keypoints are detected by looking for extrema, both in space and scale, in $\mathsf{D}(x, y, \sigma)$. The scale σ of detection is kept for subsequent normalization of the description patch in order to assure scale invariance.

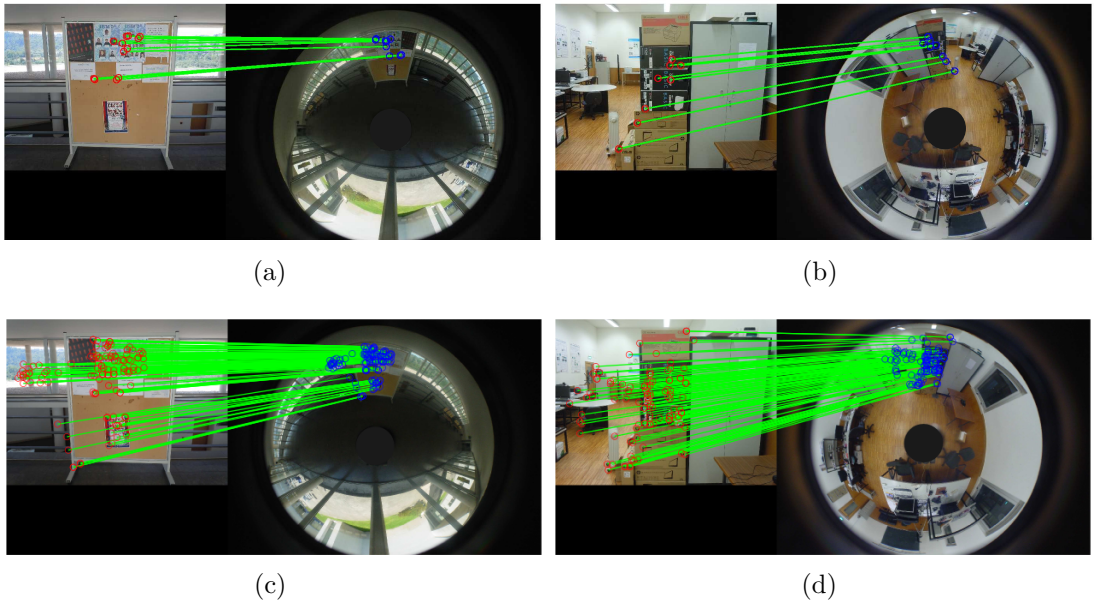
- **Keypoint description:**

  The description is performed by considering a local image neighborhood around the detected point. To each keypoint is assigned one or more orientations based on local image gradient directions. Representing the keypoint descriptor relative to this orientation allows to achieve invariance to image rotation. Then, a set of orientation histograms are created on 4x4 pixel neighbors with 8 bins each. These histograms are computed from gradient magnitude and orientation values of samples in a 16x16 region around the keypoint. The magnitudes are further weighted by a Gaussian function.

2.1.2 Difficulties in establishing matches between perspective images and paracatadioptric images

In order to evaluate the ability of the SIFT algorithm in matching perspective images with catadioptric views we ran the standard Lowe implementation in Figure 2.1(a) and Figure 2.1(b). It can be observed that there are very few correct matches. This is not surprising because a careful observation shows that, due to the specific image formation process in the catadioptric system, the local patches around keypoints are reflected when compared to the perspective. In order to solve the above issue we reflected the omnidirectional view before applying standard SIFT. The improvements can be observed in Figure 2.1(c) and Figure 2.1(d). The reflection improved the results, but further improvements can be achieved by taking into account the image geometry. When using omnidirectional images, their particular geometry causes partial scale changes in different regions of the image (e.g. paracatadioptric images are sampled more densely in the outer parts than in the center). Classical scale-invariant feature detection algorithms like SIFT do not

take into account these particular characteristics of the images. In order to overcome this problem, alternative approaches are presented.



(a) (b)

(c) (d)

Figure 2.1: Examples of feature matching between perspective and catadioptric images using SIFT algorithm. In Figure 2.1(a) and Figure 2.1(b) we ran the standard Lowe implementation. Figure 2.1(c) and Figure 2.1(d) show the improvements when the catadioptric images are reflected before applying standard SIFT. The green lines represent correct matches.

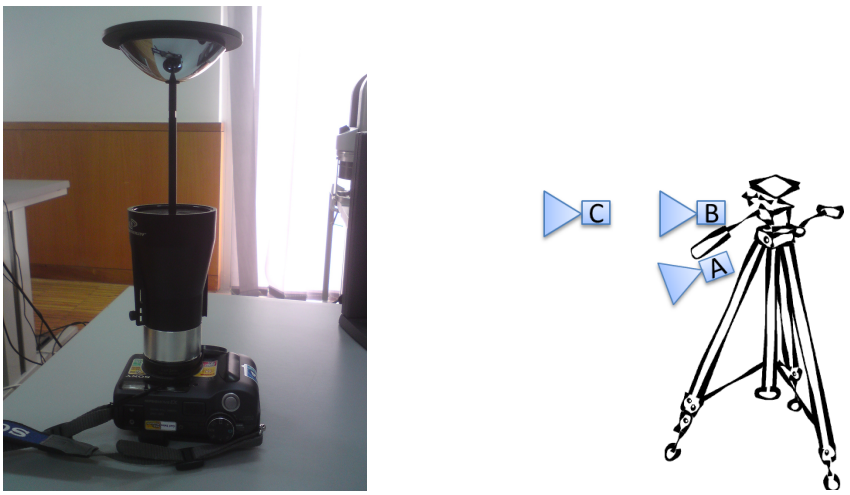
2.2 Data Set for evaluating matching performance

As stated above, this and the next chapter will discuss improvements in the SIFT matching performance in the case of mixtures of images. The design must be driven by a careful experimental evaluation. For this purpose, we collected 13 paracatadioptric images taken in different places of the campus (indoor and outdoor), using the camera of Figure 2.2(a) with a resolution of 2272x1704. Some of these images are shown in Figure 2.3(a). Perspectives of the same scene were acquired as shown in Figure 2.2(b). In order to test the different types of invariance we divided them in 4 sets:

- Set A: Taken from the same position as the paracatadioptric image and with an angle of approximately 45 degrees between the optical axis and the vertical plane (Figure 2.3(b)).

- Set B: Taken from the same position as the paracatadioptric image and with the optical axis perpendicular to the vertical plane (Figure 2.3(c)).
- Set C: Taken from a position closer to the scene, to test strong differences in scale (Figure 2.3(d)).
- Set D: Taken from different positions and view points relatively to the paracatadioptric images, to test strong viewpoint changes (Figure 2.3(e)).

The resolution of the perspective images is 1600x1200. All the images were acquired with the auto-exposure and the auto-focus modes turned on. The camera settings were maintained for the entire procedure.



(a) (b)

Figure 2.2: In 2.2(a) is shown the paracatadioptric system, constituted by a standard camera (*Sony Cyber-Shot DSC-S85*) coupled with an orthographic lens and a parabolic mirror. 2.2(b) shows a representative scheme for the construction of the sets of perspective images.

2.2.1 Selection of correct matches based on geometric criteria

In order to do matching between a pair of images of the same scene, SIFT features are first extracted from both images. Each feature of an image is individually compared to the features of the other image. Correspondences are obtained based on the Euclidean distance of their feature vectors. These correspondences are identified using a typical

(a) Omnidirectional images (b) Set A (c) Set B (d) Set C (e) Set D

Figure 2.3: Examples of paracatadioptric images and the corresponding perspective images.

nearest neighbor algorithm that considers that a match is correct if the distance to the first ($d1$) and the second ($d2$) nearest neighbors verifies the following:

$$\frac{d_1}{d_2} \leq \lambda \tag{2.4}$$

with λ a pre-defined threshold. It is important to refer that the initial feature correspondence set have incorrect matches. In order to eliminate automatically the mismatches, the Random Sample Consensus (RANSAC[2]) algorithm is applied. The maximum number of inliers (correct matches) is estimated using the code provided by Peter Kovesi [14]. RANSAC allows to perform the fitting of the fundamental matrix or the homography

[2]RANSAC is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers.

from an initial set of correspondences between images, in a robust manner. This allows to analyze the total number of correct matches between views. The fundamental matrix and homography describe the projective relation between two views of the same scene: the first, through the epipolar geometry - a point in one view determines a line in the other, which is the image of the ray through that point; and the second, through homography - a point in one view determines a point in the other which is the image of the intersection of the ray with a plane. The homography is used when the scene is planar, or there are very small depth changes. The fundamental matrix is applied in the remaining cases.

2.3 Geometric Considerations for Image Warping

A possible solution to correct the radial distortion is to warp the paracatadioptric images, approximating them by a pin-hole projection. In order to understand how to do the warping, the catadioptric image formation process will be explained based on [5, 6].

As referred in [5], 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 $\mathbf{X_h}$ (points in the world reference frame) are mapped into projective rays/points $\mathbf{x}$ in the catadioptric system reference frame (the mirror reference frame) centered in the effective view point (these projective rays join the world point with the effective view point of the paracatadioptric system). The transformation is linear and can be described by a 3 x 4 matrix P such that

$$\mathbf{x} = \mathsf{P}\mathbf{X_h} \tag{2.5}$$

   where $\mathsf{P} = \mathsf{R_c}\,[\,\mathsf{I}\,|\,-\,\mathsf{C}\,]$ transforms points in the world reference frame to projective rays in the catadioptric system reference frame (C represents the world origin coordinates in the mirror reference frame, $\mathsf{R_c}$ is the rotation matrix between the two coordinate systems and I is a 3 x 3 identity matrix).

2. The non-linear function h maps points $\mathbf{x}$ into points $\bar{\mathbf{x}}$ in a second oriented projective plane.

$$h(\mathbf{x}) = (x, y, z + \sqrt{x^2 + y^2 + z^2}) \tag{2.6}$$

3. Projective points $\hat{\mathbf{x}}$ in the catadioptric image plane are obtained after a transformation described by $\mathsf{H_c}$

$$\hat{\mathbf{x}} = \mathsf{K_c} \underbrace{\begin{bmatrix} 2p & 0 & 0 \\ 0 & 2p & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathsf{H_c}} \bar{\mathbf{x}} \tag{2.7}$$

where $\mathsf{H_c}$ depends on the mirror parameters (lactus rectum of the parabolic mirror p) and camera intrinsic parameters $\mathsf{K_c}$.

Figure 2.4 presents a schematic to understand this model with more detail. The coordinate system with origin $\mathbf{O}$ in the effective view point is the mirror reference frame and the sphere is centered in $\mathbf{O}$ and has unit radius. The oriented projective ray $\mathbf{x}$ intersects the unit sphere in the point $\mathbf{X_c}$. Considering a point $\mathbf{O_c}$ with coordinates $(0,0,-\xi)$ in the mirror reference frame, to each $\mathbf{x}$ corresponds an oriented projective ray $\bar{\mathbf{x}}$ joining $\mathbf{O_c}$ with $\mathbf{X_c}$. The non-linear mapping $h()$ corresponds to projecting the scene in the unit sphere surface and then re-projecting the points on the sphere into a plane from a novel projection center ($\mathbf{O_c}$). Points in the image plane $\hat{\mathbf{x}}$ are obtained after a transformation $\mathsf{H_c}$ of 2D projective points $\bar{\mathbf{x}}$. For the parabolic case, $\xi = 1$.

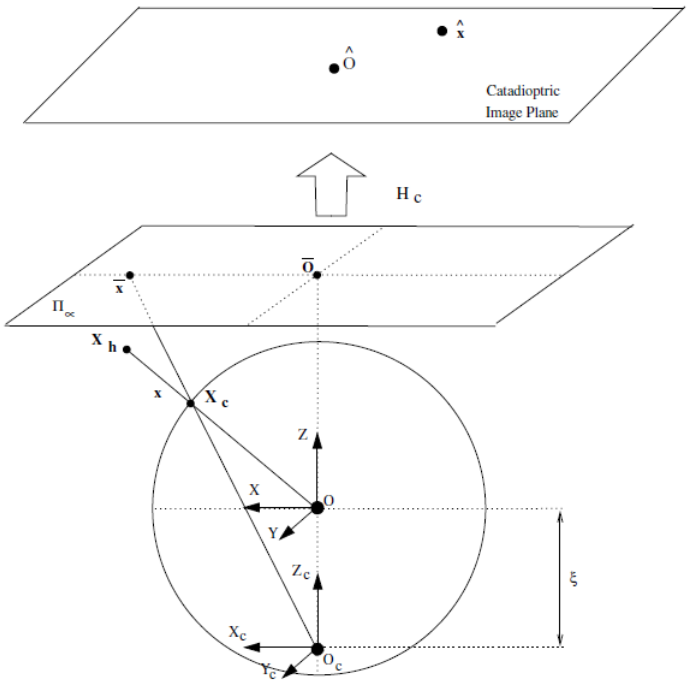


Figure 2.4: The sphere model for central catadioptric imaging [5].

2.3.1 Virtual Camera Planes

As explained before, to each point $\hat{\mathbf{x}}$ in the catadioptric image plane corresponds one projective ray $\mathbf{x}$. This projective ray can be obtained by $\mathbf{x} = h^{-1}(\mathsf{H}_c^{-1}\hat{\mathbf{x}})$, with

$$h^{-1}(\mathbf{x}) = (2xz, 2yz, z^2 - x^2 - y^2)^T \tag{2.8}$$

Supposing that the projection center of a virtual perspective camera is coincident with the effective viewpoint of the paracatadioptric system, the mathematical relation between a point in the perspective image $\mathbf{x_p}$ (rectified image) and a projective ray $\mathbf{x}$ is $\mathbf{x_p} = \mathsf{K}\,\mathsf{R}\,\mathbf{x}$ where K is the matrix of intrinsic parameters, and R is the rotation matrix between the reference frame attached to the virtual camera and the sensor coordinate system [5].

Considering $\mathsf{R} = \mathsf{I}$, the mapping between the virtual perspective image and the catadioptric image can be described by

$$\mathbf{x_p} = \mathsf{K}\,h^{-1}(\mathsf{H}_c^{-1}\,\hat{\mathbf{x}}) \tag{2.9}$$

Points $\hat{\mathbf{x}}$ in the catadioptric image are mapped into points $\mathbf{x_p}$ in the rectified image.

In practice, the rectified image is generated using the inverse of the mapping provided in Equation 2.9. All points in the perspective image are mapped into points in the catadioptric image, and the brightness of the points in the perspective image is computed using bilinear interpolation. The equation used is then

$$\hat{\mathbf{x}} = \mathsf{H}_c\,h(\mathsf{K}^{-1}\,\mathbf{x_p}) \tag{2.10}$$

which corresponds to the inverse of Equation 2.9. This approach was also taken in the other rectification methods tested in the thesis. In Figure 2.5(b) an example of a rectified image using this method is showed. As it can be seen, the rectified images obtained using this mapping have great differences in viewpoint and rotation in relation to the perspective images of Figure 2.3. Therefore, to obtain rectified views close to the perspective images, matrix R must also be considered:

$$\mathbf{x_p} = \mathsf{K}\,\mathsf{R}\,h^{-1}(\mathsf{H}_c^{-1}\,\hat{\mathbf{x}}) \tag{2.11}$$

where R is the rotation matrix that allows for different "rectified views". Four different rotation matrices corresponding to the orientations 0°, 90°, 180° and 270° were chosen, obtaining four different perspective images from the paracatdioptric image. Figure 2.5(c) shows the perspective images obtained from the paracatadioptric image of Figure 2.5(a) using this method.

(a) Original image (b) Perspective (down view)

(c) Perspective (lateral views)

(d) Cylinder

Figure 2.5: Images obtained from the warping of paracatadioptric image of Figure 2.5(a).

2.3.2 Cylindrical coordinates

Instead of back-projecting the omnidirectional images into planes to obtain virtual perspectives, we can map the original image into a cylinder and unfold it to obtain a panorama. If $\mathbf{x}$ is the back-projected point

$$\mathbf{x} = (x, y, z)^T = h^{-1}(\mathsf{H_c}^{-1}\,\hat{\mathbf{x}}) \tag{2.12}$$

then the cylindrical coordinates are

$$\begin{cases} \theta = s \cdot arctg\left(\dfrac{x}{y}\right) \\ h = s \cdot \dfrac{z}{\sqrt{x^2 + y^2}} \end{cases} \tag{2.13}$$

where s is a scaling factor (the *radius* of the cylinder) and has been set to $s = f$ to minimize the distortion (scaling) near the center of the image, where f is the focal length [30]. In Figure 2.5(d) is showed the paracatadioptric image of Figure 2.5(a) after unwarping to cylindrical coordinates.

2.3.3 Calibration Matrix

As seen before, in order to unwarp the catadioptric images, the calibration matrix $\mathsf{H_c}$ (Equation 2.14) must be known. One way to obtain $\mathsf{H_c}$ is to use the *CatPack toolbox* [1] developed by João Pedro Barreto which is a Matlab software package for the calibration of Central Catadioptric Cameras using line images.

The matrix $\mathsf{H_c}$ can also be roughly estimated when the vertical FOV is known. As explained before, to a point $\hat{\mathbf{x}}$ in the catadioptric image plane corresponds one projective ray $\mathbf{x}$, obtained with $\mathbf{x} = h^{-1}(\mathsf{H_c}^{-1}\,\hat{\mathbf{x}})$.

$$\mathsf{H_c} = \begin{bmatrix} f_x & 0 & C_x \\ 0 & f_y & C_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.14}$$

If we assume unitary aspect ratio $f = f_x = f_y$, and the center (C_x, C_y) coincident with

the center of the circular image region, it comes that

$$\mathsf{H_c} = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.15}$$

Replacing in the inverse function yields

$$\mathbf{x} = \begin{pmatrix} \frac{2\hat{x}}{f} \\ \frac{2\hat{y}}{f} \\ 1 - \frac{1}{f^2}\left(\hat{x}^2 + \hat{y}^2\right) \end{pmatrix} \tag{2.16}$$

where $\hat{x}$ and $\hat{y}$ are the non-homogeneous coordinates of the points $\hat{\mathbf{x}}$ in the catadioptric image. Considering now the point with coordinates $(r, 0)$ in the image reference frame (Figure 2.6(a)) where r corresponds to the maximum radius of the useful area of the image, its corresponding projective ray can be described by:

$$\mathbf{x} = \begin{pmatrix} \frac{2r}{f} \\ 0 \\ 1 - \frac{r^2}{f^2} \end{pmatrix} = \begin{pmatrix} \frac{2rf}{f^2 - r^2} \\ 0 \\ 1 \end{pmatrix}$$

Therefore, from Figure 2.6(b), if the projective ray $\mathbf{x}$ corresponds to the point with coordinates $(r, 0)$, and θ corresponds to the vertical FOV, it follows

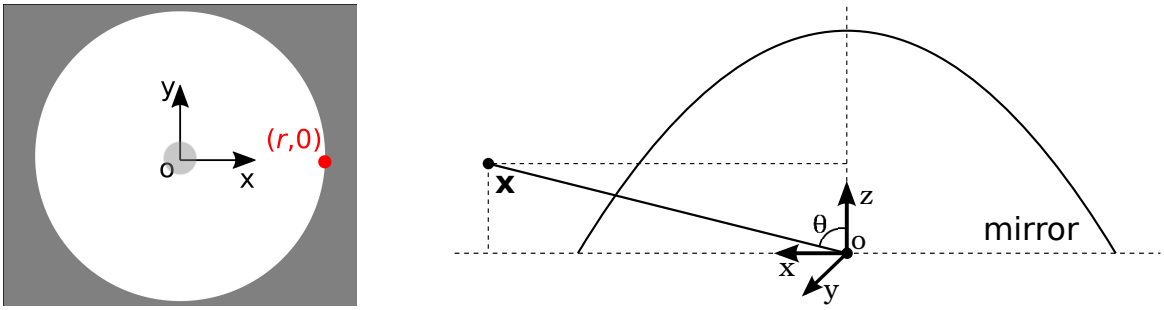
$$tg(\theta) = \frac{2rf}{f^2 - r^2}$$

Assuming that θ is known and is such that $\frac{\pi}{2} < \theta < \pi$, we are able to estimate the focal length:

$$f = \frac{r}{tg(\theta)}\left(1 - \sqrt{1 + tg(\theta)^2}\right) \tag{2.17}$$

Using this expression, $\mathsf{H_c}$ can be computed assuming the field of view is known.

One simple practical procedure to estimate the vertical FOV θ of the paracatadioptric system is shown in Figure 2.7 and it can be obtained from Equation 2.18 (this value corresponds to half the real vertical FOV of the paracatadiotric image).



(a) (b)

Figure 2.6: (a) Schematic of a catadioptric image with a point with coordinates $(r, 0)$ marked in red. (b) Representation of a projective ray $\mathbf{x}$ in the mirror reference frame.

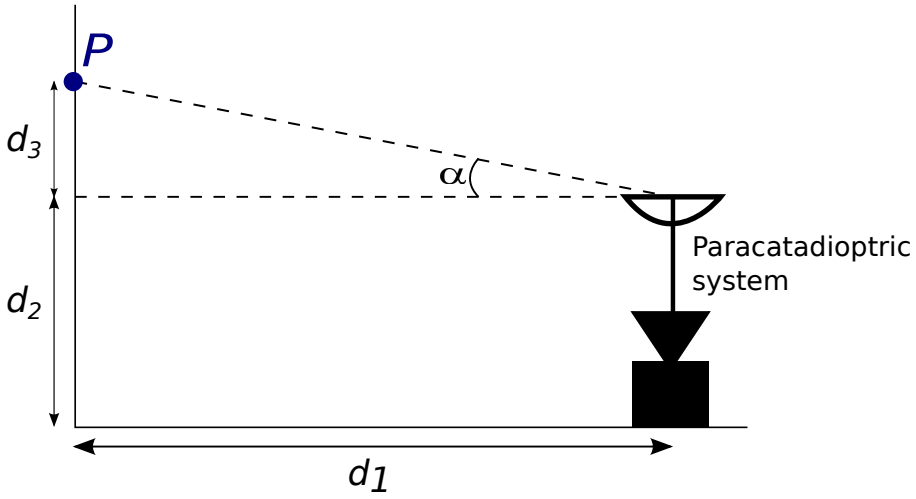


Figure 2.7: Schematic of the procedure to estimate the vertical FOV of the paracatadioptric system. Suppose that the horizontal plane where the paracatadioptric system is laying is a table and the vertical plane (where distances d_2 and d_3 are measured) is a wall. d_1 is the distance between the wall and the paracatadioptric system and d_2 is the height of the paracatadioptric system. Distance d_3 can be measured by locating the highest point P of the wall that is visible in the paracatadioptric image.

$$\theta = 90^o + \arctan\left(\frac{d_3}{d_1}\right) = 90^o + \alpha \tag{2.18}$$

2.3.4 Dealing with the Resolution

In order to coherently compare the different rectification methods, the resolutions of the warped images must be consistent. Considering a catadioptric image, its useful area can be limited by two circumferences with radius r_i and r_e as represented in Figure 2.8. r_i and r_e are the minimum and maximum radius of the catadioptric image, respectively.

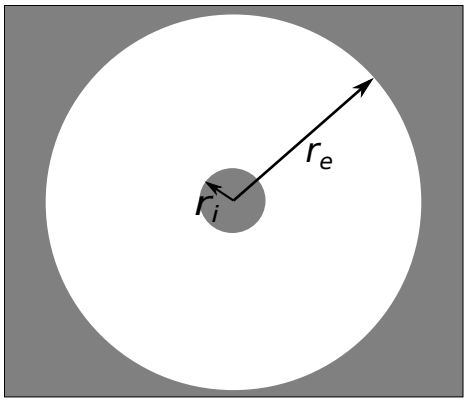


Figure 2.8: Maximum (r_e) and minimum (r_i) radius of the catadioptric image, limiting the useful area.

In this thesis, the values of r_i and r_e were obtained using the *Boundary Detector* developed in [21]. The resolution for the cylindrical images (Section 2.3.2) was computed using Equations 2.19. These equations allow to obtain approximately the same number of pixels in the areas with useful information for both catadioptric and cylindrical images. The resolution obtained was 3890x1489. In the *virtual camera planes* method (Section 2.3.1), the rectified images have a FOV in the horizontal plane of 108^o. Therefore, a consistent resolution for this case can be obtained by dividing the horizontal resolution of the cylindrical image by 3.33 ($360^o \div 108^o$) obtaining 1168x1489.

$$width = 2\,(r_e - r_i) \quad height = \frac{2\,\pi\,(r_e + r_e/2\,)}{2} \tag{2.19}$$

2.4 Experimental Results

This section analyses feature matching between the different unwarped images (Section 2.3) from a set of 13 paracatadioptric images and the corresponding perspective images (see Appendix A). For the case described in Section 2.3.1, the image corresponding to the scene of the perspective image (same orientation) is used. Matching using a rectified image with an offset of 45^o to the correct orientation is also tested. Additionally, matching using the original SIFT algorithm on the reflected paracatadioptric images was done (*Sift*). The results are summarized in Table 2.1 and in Figure 2.9 (extensive results are reported in Appendix B). Figure 2.10 shows the results of feature detection and matching between

the rectified views from paracatadioptric image of Figure A.1(c) and the corresponding perspective image of Set B.

	Inliers/Matches			
	Sift	**VCP with offset**	**VCP aligned**	**Cylinder**
Set A	97.31/305.23 (29.47%)	53/276 (19.03%)	93.77/319.15 (27.91%)	118.69/334.08 (31.70%)
Set B	131.77/389.77 (32.00%)	61.85/345.23 (17.33%)	122/393 (29.96%)	146.31/407.31 (34.43%)
Set C	62.62/463.38 (10.81%)	29/453.23 (6.08%)	52.23/478.92 (9.77%)	71.08/549.77 (11.64%)
Set D	43.23/260.85 (15.93%)	26.69/263.23 (9.83%)	45.92/273 (17.05%)	58.15/277.69 (19.55%)

Table 2.1: Results of feature matching between the rectified views (virtual camera planes (VCP) with offset, VCP aligned and cylindrical images) from a set of 13 paracatadioptric images and the corresponding perspective images. Matching results using the original SIFT algorithm on the reflected paracatadioptric images are also presented (*Sift*). For each case, the average values for the number of inliers (correct matches), total number of matches and inliers percentage are represented. These values were computed from Table B.1.

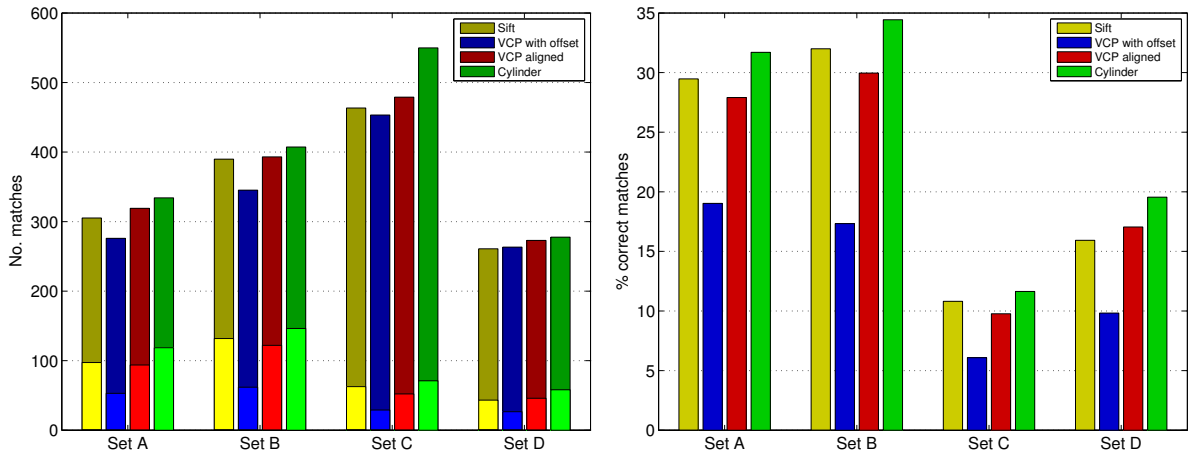
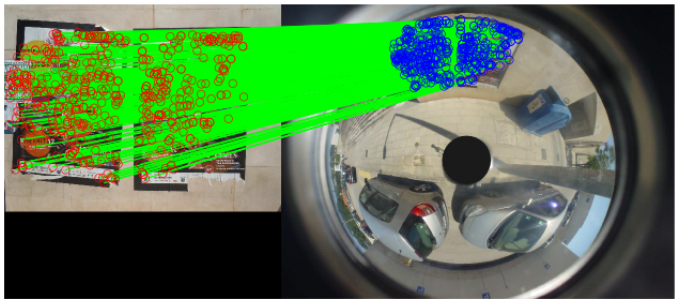
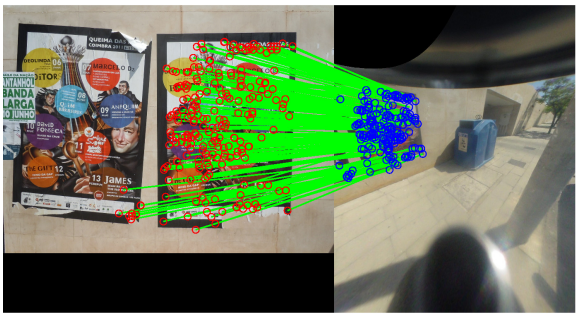


Figure 2.9: Average values for the number of matches (a) and percentage of correct matches (b) in the different cases obtained from Table B.1. In (a), the fraction of the bars with lighter color correspond to the average number of correct matches.

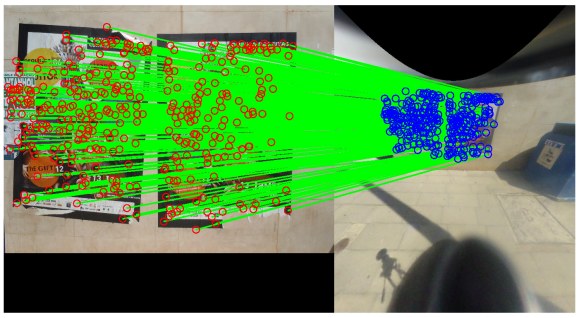
From the results obtained, it can be concluded that the most suitable method is rectifying the paracatadioptric images to cylindrical coordinates. It offers good results, outperforming VCP aligned and VCP with offset in 73.1% and 92.3% of the cases, respectively. Also, it outperforms *Sift* in 75% of the cases. The criteria for comparing the different methods is the number of correct matches (inliers). The higher the number of inliers the better the performance of a given method. In cases where the same number of



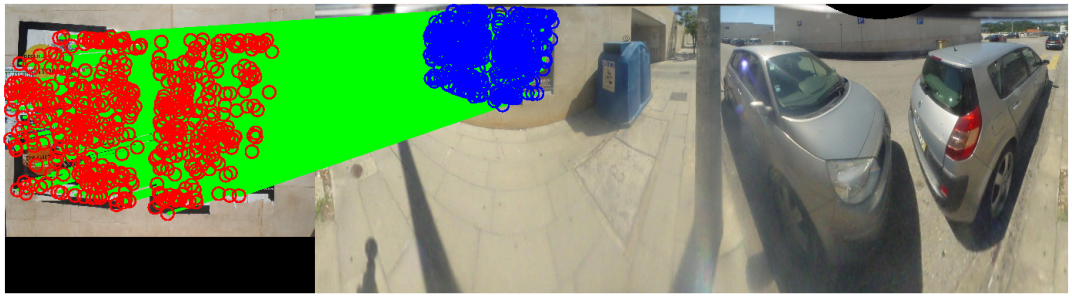
(a) Sift



(b) VCP with offset



(c) VCP aligned



(d) Cylinder

Figure 2.10: Results of feature matching between the rectified views (virtual camera planes (VCP) with offset (2.10(b)), VCP aligned (2.10(c)) and cylindrical images (2.10(d)) from paracatadioptric image of Figure A.1(c) and the corresponding perspective image of Set B. Results using the original SIFT algorithm on the reflected paracatadioptric images are also presented (2.10(a)). The green lines represent correct matches.

inliers is obtained, the inliers percentage is used as performance measure. As it can be seen, when the rotation of the rectified image does not match with the perspective image (VCP with offset) the results are poor. Also, it is quite intriguing that SIFT applied over the catadioptric images outperforms *VCP aligned* (with the exception of set D). This fact is explained by the lower resolution obtained in the VCP images, when compared with the cylinder and the catadioptric images. It was assumed a field-of-view of 108 degrees for the perspectives, which resulted in a loss of detail in the *useful* region of the image.

Additionally, analyzing Table B.1, the results obtained with Set C were inferior to Set B in all cases. This occurred because the differences in scale between images from Set C and the paracatadioptric/rectified images are greater than with Set B. The poorest results occurred with set D. This is due to the great differences in viewpoint between the perspective images from this set and the paracatadioptric/rectified images. Finally, Figure 2.9 shows that both the average number of matches and percentage of correct matches are higher in the cylinder method for all perspective images sets.

Chapter 3

Improving Hybrid Image Matching

This chapter proposes several modifications to the original SIFT algorithm in order to obtain better results of matching between perspective and paracatadioptric images (or rectified views from the paracatadioptric images). Based on the experimental results, a change in the initial scale for the Gaussian blurring in SIFT algorithm is proposed, in order to eliminate higher frequency features in the perspective images, which were generating false matches. As an additional unwarping alternative, the transformation to polar coordinates used by Luis Puig *et al.* [15] is also explained and tested. Finally, a new approach for image adaptive blurring that takes into account the geometry of the paracatadioptric images is introduced based on [17]. This approach adds low complexity to original SIFT and processes the images directly in the plane.

3.1 Resolution issues

SIFT algorithm detects features in the scale-space obtained by low-pass filtering using a variable-scale Gaussian function. This allows to detect features at different scales. It was observed that SIFT gives many false matches when matching perspective with omnidirectional images. Most errors are due to matching a high resolution feature in the perspective image with a low resolution feature in the omnidirectional image. Therefore, a solution to overcome this problem is increasing the initial scale in the construction of the scale-space of the SIFT algorithm for the perspective images. The initial octaves are no longer considered (higher frequencies), leading to an improving of the matching.

The number of octaves to neglect can be obtained experimentally. Latter in this chapter results will be presented showing the improvements in feature matching due to this change in the initial scale of the SIFT algorithm.

3.2 Transformation to Polar Coordinates (*Polar*)

Another approach to do matching between images coming from central catadioptric systems and conventional cameras was developed by Luis Puig *et al.* [15]. As showed before, if SIFT features extracted in an omnidirectional image are matched to features extracted in a perspective image, the results are poor. In [15], the omnidirectional images are warped using a transformation to polar coordinates (see Equations 3.1). While in Section 2.3.2 (transformation to cylindrical coordinates), the calibration matrix must be known, this transformation does not require camera calibration. For image matching, they first extract SIFT features from the perspective and warped images to establish pairs of putative corresponding points between the views.

$$\theta = arctg\left(\frac{\hat{y}}{\hat{x}}\right) \quad \rho = \sqrt{\hat{x}^2 + \hat{y}^2} \tag{3.1}$$

3.3 Implicit filtering (*CylSIFT*)

This section introduces a new approach for image adaptive blurring that takes into account the geometry of the paracatadioptric images.

3.3.1 Keypoint Detection

The objective here is to generate a scale-space representation equivalent to the one that would be obtained by filtering the paracatadioptric image in the absence of distortion. To achieve such goal and based on [17], the distortion correction will be performed in an implicit manner, by adapting the convolution kernel that is used directly over the paracatadioptric image.

Equations 2.12 and 2.13 describe the mapping from the paracatadioptric image to a

"rectified view" in cylindrical coordinates. Considering the mapping function described by Equation 2.16, Equations 2.13 become:

$$f_u(x,y) = f \cdot arctg\Big(\frac{x}{y}\Big) \tag{3.2}$$

$$f_v(x,y) = \frac{f}{2} \cdot \Big(\frac{f}{\sqrt{x^2+y^2}} - \frac{\sqrt{x^2+y^2}}{f}\Big) \tag{3.3}$$

Let G_σ be a bi-dimensional Gaussian function with standard deviation σ, $\hat{\mathtt{I}}$ the undistorted image, and $\mathtt{I}$ the distorted image. The value of the blurred undistorted image $\hat{\mathsf{L}}_\sigma$ at pixel (s,t) is given by

$$\hat{\mathsf{L}}_\sigma(s,t) = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} \hat{\mathtt{I}}(u,v)\, \mathsf{G}_\sigma(s-u, t-v) \tag{3.4}$$

This is the convolution that SIFT performs for the case of the image being rectified for correcting the distortion. However, and since the objective is to work directly with the distorted image $\mathtt{I}$, the undistorted image $\hat{\mathtt{I}}$ can be replaced by its distorted counterpart, taking into account the mapping functions 3.2 and 3.3. Considering that

$$\hat{\mathtt{I}}(u,v) = \mathtt{I}(f_u(x,y), f_v(x,y)), \tag{3.5}$$

and changing the variables (u,v) by (x,y) in Equation (3.4), it arises:

$$\begin{aligned}\hat{\mathsf{L}}_\sigma(s,t) = \sum_{x=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \sum_{y=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \mathtt{I}(x,y)\, \mathsf{G}_\sigma(s - f_u(x,y),\\ t - f_v(x,y))\end{aligned} \tag{3.6}$$

Since L_σ is the distorted version of the smoothed image $\hat{\mathsf{L}}_\sigma$, same reasoning is applied changing the undistorted coordinates (s,t) by their distorted counterparts (h,k). It follows that

$$\begin{aligned}\mathsf{L}_\sigma(h,k) = \sum_{x=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \sum_{y=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \mathtt{I}(x,y)\, \mathsf{G}_\sigma(f_u(h,k) - f_u(x,y),\\ f_v(h,k) - f_v(x,y)),\end{aligned} \tag{3.7}$$

Then, using Equation 3.2 and Equation 3.3 in Equation 3.7 and after some algebraic manipulation one can obtain:

$$\mathtt{L}_\sigma(h,k) = \sum_{x=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \sum_{y=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \mathtt{I}(x,y)\, \mathsf{G}_\sigma\Big(f \cdot \arctan\Big(\frac{h-c_x}{k-c_y}\Big) - f \cdot \arctan\Big(\frac{x-c_x}{y-c_y}\Big), \frac{f}{2}\Big(\frac{f}{\sqrt{r_{hk}}} - \frac{\sqrt{r_{hk}}}{f} - \frac{f}{\sqrt{r_{xy}}} + \frac{\sqrt{r_{xy}}}{f}\Big)\Big) \tag{3.8}$$

where $r_{hk} = (h-c_x)^2 + (k-c_y)^2$ and $r_{xy} = (x-c_x)^2 + (y-c_y)^2$. The focal length f is obtained with Equation 2.17. Note that now the smoothing kernel depends on (x,y) and (h,k) and it is no longer a straightforward Gaussian convolution. For each radius, the adaptive blurring kernel has the same shape, but with different orientations (see Figure 3.1).

The standard two-dimensional Gaussian $\mathsf{G}(\cdot;\sigma)$ is a rank 1 matrix that can be written as the outer product of two one-dimensional Gaussian of the the same standard deviation σ:

$$\mathsf{G}(\cdot;\sigma) = \mathbf{g}_y(\cdot;\sigma)\mathbf{g}_x(\cdot;\sigma)$$

where $\mathbf{g}_x(\cdot;\sigma)$ a row vector and $\mathbf{g}_y(\cdot;\sigma)$ is a column vector. Due to the separability property of the Gaussian filter, the standard scale-space image representation can be computed in a computation affordable manner by:

$$\mathtt{I} * \mathsf{G}(\cdot;\sigma) = \mathbf{g}_x(\cdot;\sigma) * (\mathtt{I} * \mathbf{g}_y(\cdot;\sigma))$$

Instead of computing the *cylindrical* Gaussian for each image pixel position, we approximate the Equation 3.8 by the best rank 1 Gaussian filter obtained through Singular Value Decomposition.

$$\Big[\mathsf{U} \quad \mathsf{S} \quad \mathsf{V}\Big] = \mathtt{SVD}(\mathsf{G}_\sigma) \tag{3.9}$$

Now we get the horizontal and vertical vectors from the first columns of U and V,

$$\mathbf{g}_v = \mathsf{U}(:,1) \times \mathsf{S}(1,1) \qquad \mathbf{g}_h = \mathsf{V}^\mathsf{T}(:,1) \times \mathsf{S}(1,1) \tag{3.10}$$

that permit to obtain the rank 1 Gaussian filter that better approximates Equation 3.8

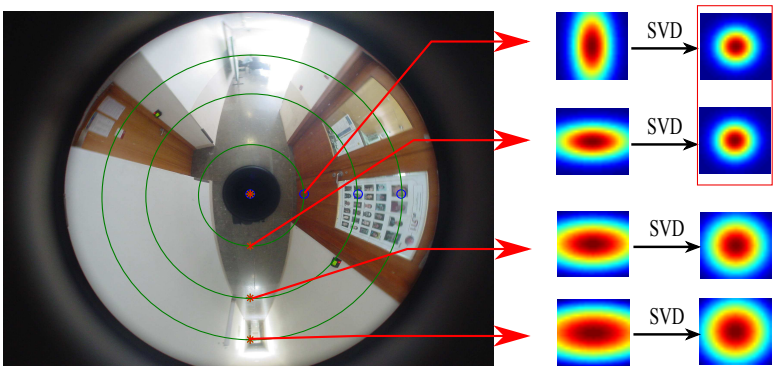


Figure 3.1: *Cylindrical* Gaussian filters decomposition. The filters support region increases from the center towards the periphery to adapt to the non-uniform sampling of the catadioptric images. For each image radius, the gaussian filters have the same shape, but different orientations.

by:

$$\mathsf{G}_{\sigma,rank=1} = \mathbf{g}_v * \mathbf{g}_h; \tag{3.11}$$

The computational advantages of this decomposition are twofold:

- For every image radius the Gaussian filter is separable, which permits a considerable speedup of the convolution process;
- A filter bank is computed off-line and then is loaded into memory. The same filter bank is used for all images of the dataset.

3.3.2 Keypoint Description

In order to minimize the radial distortion effect, the Gaussian weighting function used to assign the gradient magnitudes to the descriptor, is changed. In the standard SIFT descriptor the gradient weighting is proportional to the scale of selection of the keypoints. However, the distortion effect becomes more pronounced near the keypoint patch periphery, precluding successful matches to be established. To compensate this effect, we propose a simple yet effective way of dealing with the non-linear changes on the gradient magnitudes. Instead of using the scale of selection of the keypoints, we change the Gaussian weighting function to be $\frac{1}{2}\sigma$, being sigma the scale of selection of the keypoints. This factor was experimentally selected and provides the best results.

3.4 Performance evaluation

In this section, different methods for feature extraction and description for paracatadioptric images are evaluated. In this hybrid matching comparison, SIFT is used to extract features in the perspective images. For all methods, the initial scale for the construction of the scale-space in SIFT algorithm is doubled (from 1.6 to 3.2) for the perspective images. The difference relies on the method used to extract features in the catadioptric/rectified views. The following approaches are considered:

- Application of SIFT over paracatadioptric images with the reflection of the paracatadioptric images (*Sift-LF*).
- Application of SIFT over the rectified paracatadioptric images to polar coordinates, as explained in Section 3.2 (*Polar*).
- Application of SIFT over the rectified paracatadioptric images to cylindrical coordinates (*Cylinder-LF*).
- Application of SIFT with implicit filtering (*CylSIFT*) over the paracatadioptric images, as explained in Section 3.3.

The same sets of images as in Chapter 2 are used. Matching results are summarized in Table 3.1 and in Figure 3.2 (extensive results are reported in Appendix B). Figure 3.3 shows the results of feature matching between the paracatadioptric/rectified image of Figure A.1(k) and the corresponding perspective image of set B.

	Inliers/Matches			
	Sift-LF	*Polar*	*Cylinder-LF*	*CylSIFT*
Set A	105.46/202.54 (50.09%)	108.54/197.38 (52.08%)	128.38/217.31 (56.45%)	138.15/203.77 (65.68%)
Set B	141/253.31 (54.24%)	138.54/248.54 (54.42%)	158.08/262.46 (59.35%)	188.62/271.92 (70.55%)
Set C	89.62/247.92 (34.22%)	90/242 (35.63%)	105.23/260 (38.53%)	126.92/229.62 (51.68%)
Set D	42.31/145.69 (27.63%)	47.77/172.62 (29.70%)	56.46/174.23 (34.32%)	65.69/123.08 (48.22%)

Table 3.1: Results of feature matching between the perspective and the paracatadioptric/rectified images using different approaches (*Sift-LF*, *Polar*, *Cylinder-LF* and *CylSIFT*). For each case, the average values for the number of inliers (correct matches), total number of matches and inliers percentage are represented. Detailed results are reported in Appendix B.

From the results obtained the following conclusions can be taken:

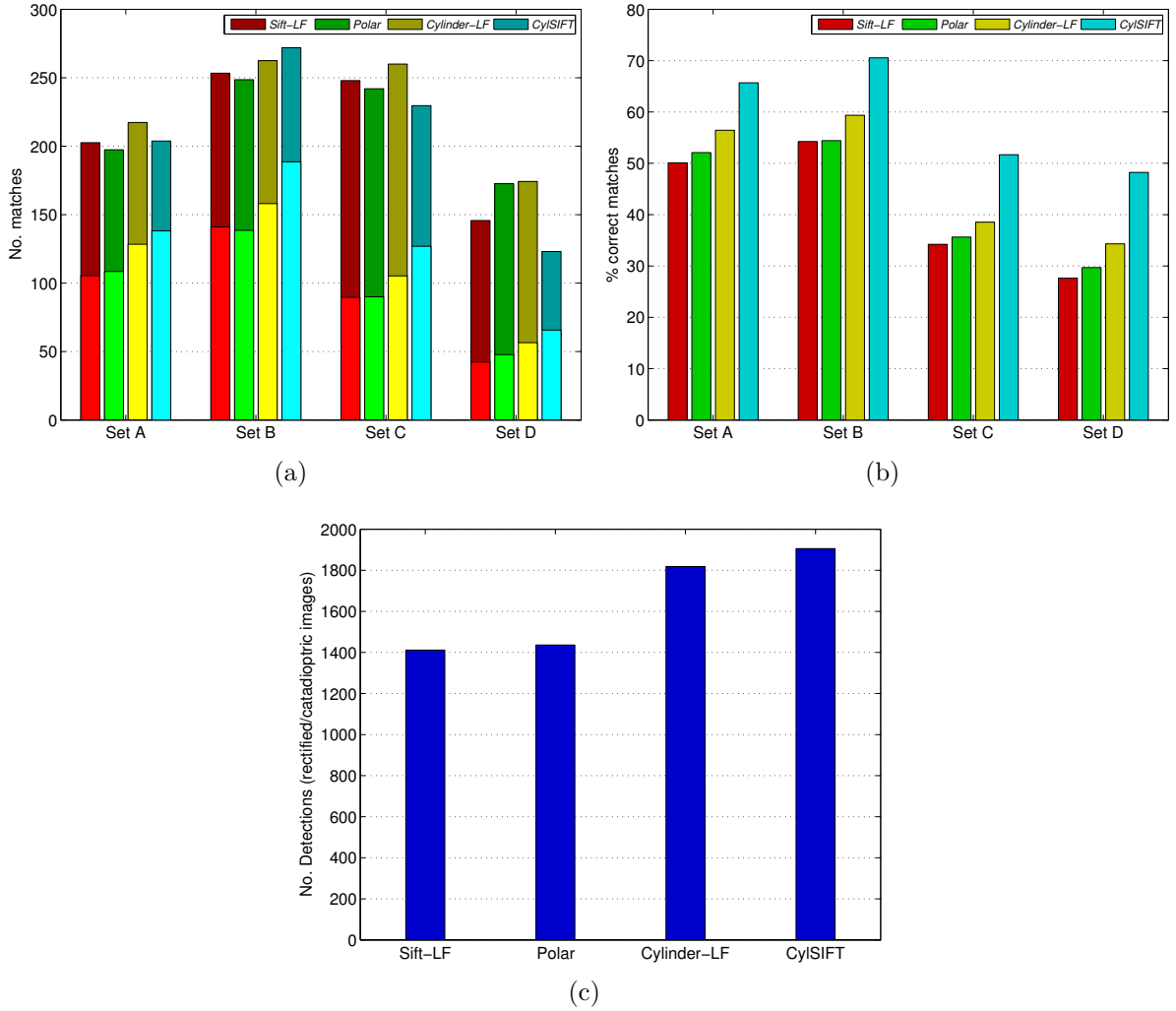
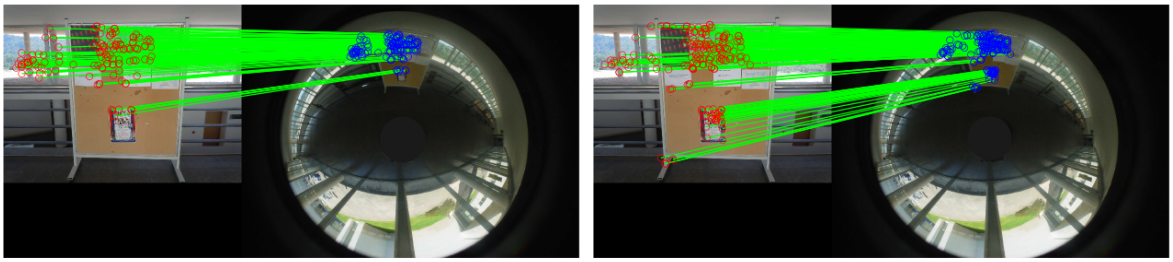


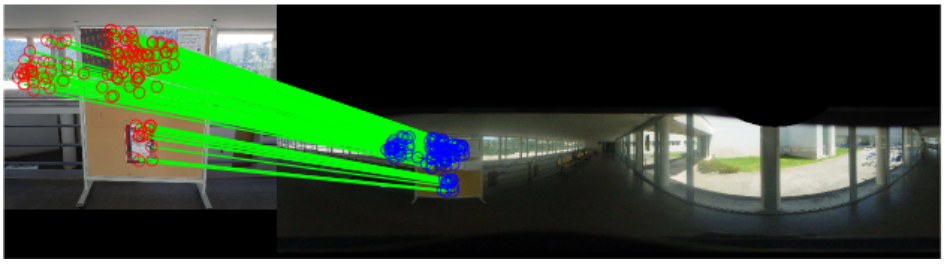
Figure 3.2: Average values for the number of matches (a), percentage of correct matches (b) and number of detections (c) in the different cases obtained from Tables B.3 and B.4. In (a), the fraction of the bars with lighter color correspond to the average number of correct matches.

- Doubling the initial scale in the construction of the scale-space for the perspective images, leaded to an improvement of the performance of SIFT algorithm. It can be seen that in 80.8% of the cases, *Cylinder-LF* outperformed *Cylinder*, and *Sift* was outperformed in 88.5% of the cases by *Sift-LF*. Additionally, there was a significant increase in the inliers percentage.

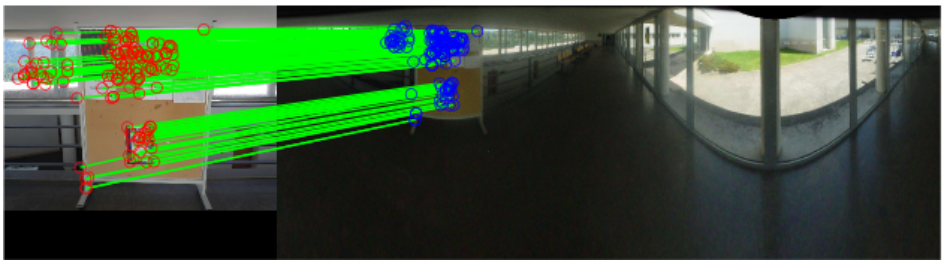
- *Cylinder-LF* outperformed *Polar* in 88.5% of the cases. The *Polar* method has the advantage of not requiring camera calibration at all. As the rectification to cylindrical coordinates uses the calibration matrix and the non-linear function characteristic



(a) *Sift-LF* (b) *CylSIFT*



(c) *Polar*



(d) *Cylinder-LF*

Figure 3.3: Results of feature matching between the paracatadioptric/rectified image of Figure A.1(k) and the corresponding perspective image of set B using different approaches (*Sift-LF*, *Polar*, *Cylinder-LF* and *CylSIFT*). The green lines represent correct matches.

of the mirror, the mapping is better justifying the results obtained.

- Finally, *CylSIFT* provided the best results outperforming *Sift-LF*, *Polar* and *Cylinder-LF* in 100%, 96.2% and 84.6% of the cases, respectively. Using *CylSIFT*, the scale-space representation of the image is obtained using adaptive filtering that compensates the distortion. This approach avoids the artifacts arising from the signal reconstruction process [9]. Additionally, it can be seen from Figure 3.2 that *CylSIFT* has the best results in terms of average number of correct matches and percentage of correct matches.

3.5 Matching Between Paracatadioptric Images

Up to now, a new method for feature detection and matching in hybrid systems was explained and compared with other competing approaches, outperforming them. As an additional result, a brief study of feature detection and matching between paracatadioptric images is performed. Once again, a set of approaches is analyzed and compared:

- Application of SIFT over the paracatadioptric images (*Sift*).
- Application of SIFT over the rectified paracatadioptric images to polar coordinates (*Polar*).
- Application of SIFT over the rectified paracatadioptric images to cylindrical coordinates (*Cylinder*).
- Application of SIFT with implicit filtering (*CylSIFT*) over the paracatadioptric images, as described in Section 3.3.
- Application of SIFT with implicit filtering over the paracatadioptric images, but in this case with a different mapping function (*cataSIFT*) as described in Section 3.5.1.

3.5.1 cataSIFT

As in Section 3.3, the objective is also to generate a scale-space representation equivalent to the one that would be obtained by filtering the paracatadioptric image in the absence of distortion. Instead of using the mapping from the paracatadioptric image to cylindrical coordinates, the mapping function described by Equation 2.9 (mapping to a "rectified view" similar to Figure 2.5(b)) is used to implicitly adapt the Gaussian filters. After some manipulation one can obtain:

$$\mathbf{x_p} = \begin{pmatrix} 2x \\ 2y \\ 1 - \frac{1}{f^2}\left(x^2 + y^2\right) \end{pmatrix} \tag{3.12}$$

and, following the same philosophy as in Section 3.3:

$$\mathtt{L}_\sigma(h,k) = \sum_{x=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \sum_{y=-\frac{1}{\sqrt{-\xi}}}^{\frac{1}{\sqrt{-\xi}}} \mathtt{I}(x,y)\, \mathtt{G}_\sigma\Big(\frac{2}{1-\gamma r^2}(h-x), \frac{2}{1-\gamma r^2}(k-y)\Big) \tag{3.13}$$

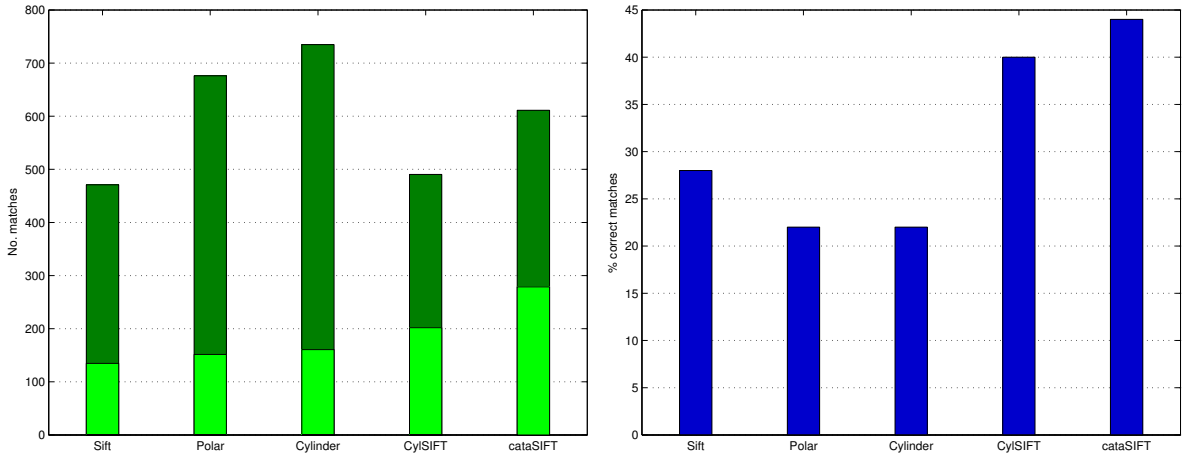
where $\gamma = \frac{1}{f^2}$ and $r = \sqrt{r_{hk}}$.

3.5.2 Results obtained

In order to compare the different methods, a set of 6 pairs of paracatadioptric images is used (see Appendix A). The results of matching between the paracatadioptric images are shown in Table 3.2 and in Figure 3.4. In Figure 3.5 are depicted the results of using *Sift* and *cataSIFT* in 2 pairs of paracatadioptric images. From the results it can be seen that *cataSIFT* outperforms all the other approaches (highest number and percentage of correct matches), being the best solution for matching paracatadioptric images.

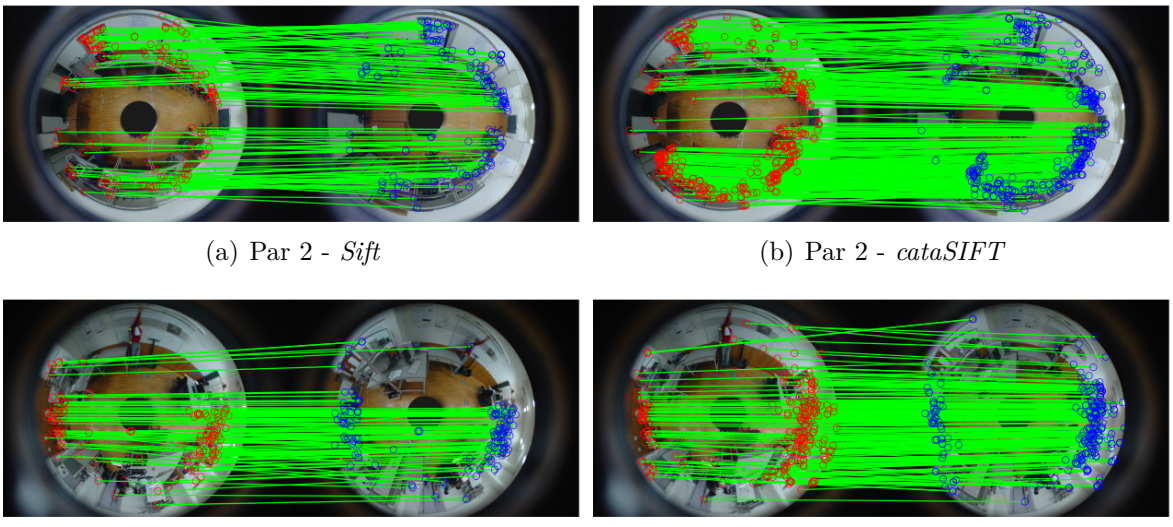
	Inliers/Matches				
	Sift	*Polar*	*Cylinder*	*CylSIFT*	*cataSIFT*
Par 1	133/420 (31.67%)	171/643 (26.59%)	175/676 (25.89%)	215/469 (45.84%)	**286/622 (45.98%)**
Par 2	147/528 (27.84%)	153/707 (21.64%)	151/841 (17.95%)	258/590 (43.73%)	**413/861 (47.97%)**
Par 3	196/539 (36.36%)	250/953 (26.23%)	259/1063 (24.37%)	300/630 (47.62%)	**328/672 (48.81%)**
Par 4	128/494 (25.91%)	137/574 (23.87%)	142/617 (23.01%)	167/399 (41.85%)	**255/508 (50.20%)**
Par 5	41/386 (10.62%)	85/507 (16.77%)	78/528 (14.77%)	72/354 (20.34%)	**94/414 (22.71%)**
Par 6	163/458 (35.59%)	113/673 (16.79%)	159/684 (23.25%)	199/500 (39.80%)	**297/590 (50.34%)**

Table 3.2: Results of feature matching in a set of pairs of paracatadioptric images (see Appendix A) using different approaches (*Sift*, *Polar*, *Cylinder*, *CylSIFT* and *cataSIFT*). For each case, the number of inliers (correct matches), total number of matches and inliers percentage is represented. The values in bold correspond to the best results for each case.



(a) (b)

Figure 3.4: Average values for the number of matches (a) and percentage of correct matches (b) in the different cases obtained from Table 3.2. In (a), the fraction of the bars with lighter color correspond to the average number of correct matches.



(a) Par 2 - *Sift* (b) Par 2 - *cataSIFT*

(c) Par 4 - *Sift* (d) Par 4 - *cataSIFT*

Figure 3.5: Examples of feature matching between paracatadioptric images. The results of *Sift* and *cataSIFT* in pairs 2 and 4 are shown. The green lines represent correct matches.

Chapter 4

Visual Place Recognition

Visual place recognition has been an active topic of research in the computer vision community [2, 3, 8, 16, 28]. It is a hard problem due to appearance variabilities arising from common image transformations, like viewpoint and scale changes. The main objective of the thesis is to perform image based localization using a database of omnidirectional images. Given an image taken from a standard camera (e.g. mobile phone), the goal is to retrieve the most similar image in the database. In the literature, very good results were shown demonstrating content based image retrieval using local scale-invariant features (e.g. SIFT) with various techniques of indexing and quantization (e.g. vocabulary trees) [25, 29]. In this chapter, the concept of vocabulary tree [25, 29] is explained, and a recognition scheme that scales efficiently to large databases of images is implemented and tested.

4.1 Recognition Using a Vocabulary of Visual Words

Sivic and Zisserman [29] perform retrieval of frames from videos using a text retrieval approach. The descriptors extracted from a set of images (e.g. using SIFT) are quantized into clusters, also called visual words. These clusters are computed using k-means algorithm performed on the descriptor vectors defining a vocabulary tree. k-means is a method of cluster analysis which aims to partition the observations into a pre-defined number of k clusters in which each observation is assigned to the cluster with the nearest mean.

4.1.1 Hierarchical *k*-means Clustering

In [25], a vocabulary tree is built using hierarchical k-means clustering. For this case, k defines the branch factor of the tree. The construction of the tree initiates by applying the k-means algorithm on the training data (image descriptors). This initial step defines k clusters where to each cluster are associated the corresponding descriptor vectors closest to its center. Then, the same process is applied recursively to each cluster, splitting it into k new groups. The vocabulary tree is built with a number of levels L, and each division into k new clusters is only defined by the distribution of the descriptor vectors that belong to the parent cluster (see Figure 4.1). The number of visual words in the vocabulary tree is then k^L. As explained in [25], building the visual vocabulary hierarchically allows to have a more efficient searching procedure and to reduce the computational cost when increasing the size of the vocabulary. In the on-line phase, at each level of the tree, the descriptor vectors are compared to the k cluster centers to choose the closest one. This is a matter of performing kL dot products (using only k-means, it would be necessary to perform k^L dot products).

Usually, increasing the size of the vocabulary lead to retrieval improvements. Although, there is always a trade-off between distinctiveness (small quantization cells and a larger vocabulary tree) and repeatability (large quantization cells).

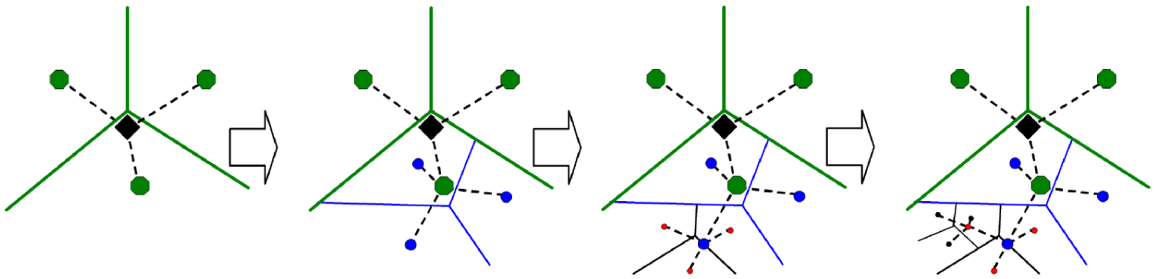


Figure 4.1: Illustration of the process of building an hierarchical vocabulary tree. The quantization is defined at each level by k centers (in this case $k = 3$) and their Voronoi regions [25].

4.1.2 Recognition Scheme

Based on this visual vocabulary, the main idea of the recognition scheme is to measure the similarity between the visual words in a query image and in the database images. This is

achieved by establishing a score for each database image, retrieving the ones with higher scores. In first place, local features are extracted from the database images. Then, each feature is quantized in the vocabulary tree, where is assigned the corresponding visual word. Therefore, each image is represented by a list of visual words instead of a set of 128-dimensional descriptor vectors which is a much more compact representation. The list of visual words from each image form a document vector whose dimension correspond to the number of visual words in the vocabulary tree (see Figure 4.2). Additionally, a weight is assigned to each visual word based on a Term Frequency - Inverse Document Frequency (TF-IDF) scheme [25],

$$w_i \;=\; \ln \frac{N}{N_i} \tag{4.1}$$

where N is the total number of images in the database and N_i the number of images in the database that contain word i. With this weighting scheme, visual words that occur in many images of the database are less discriminative and, therefore, have a low weight while visual words that occur more rarely have an higher weight. Ergo, the document vectors can be constructed by stacking the following entries

$$q_i \;=\; n_i \, w_i \tag{4.2}$$

$$d_i \;=\; m_i \, w_i \tag{4.3}$$

where n_i and m_i are the number of keypoints in the query and database images, respectively, corresponding to word i. Finally, a scoring system is built based on the normalized difference between the query and the database vectors [25]:

$$s(q,d) = \| \frac{q}{\|q\|} - \frac{d}{\|d\|} \| \tag{4.4}$$

4.1.3 Scoring System

In order to have an efficient scoring system, an inverted file is used. The inverted file has one entry for each word of the visual vocabulary. Each entry has the *ids* of the database images that contain the word and the corresponding term frequency. The scoring

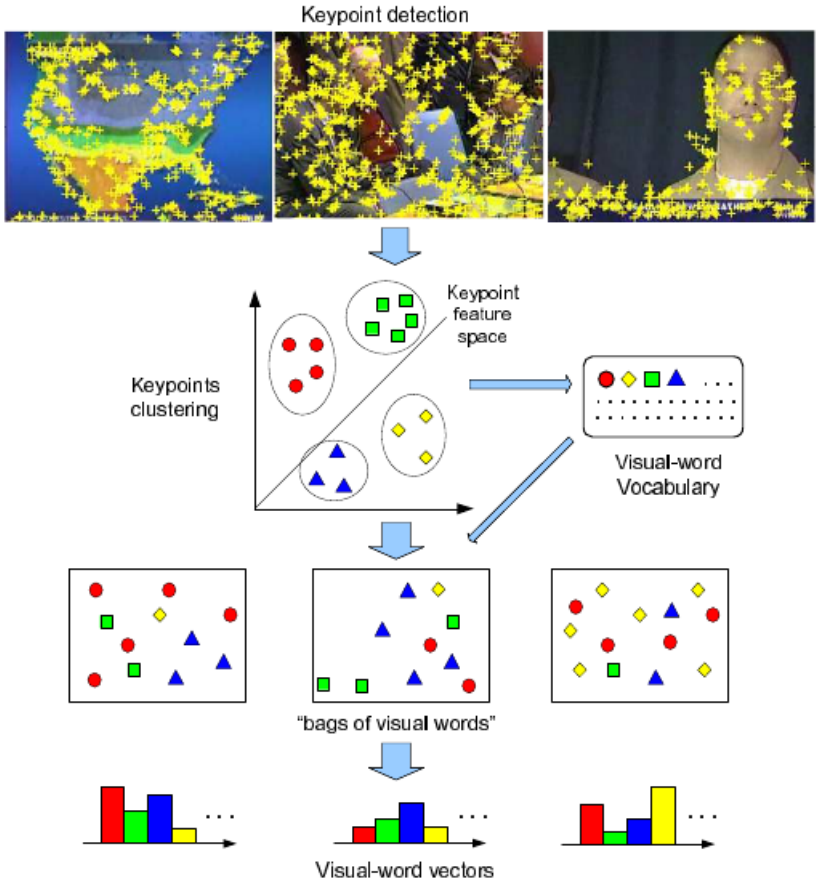


Figure 4.2: Illustration of the process of building the histograms of visual words for the database images [32].

scheme using the inverted file can be interpreted as follows [12]: Initially the scores of the database images are initialized to 0. For each visual word **i** in the query image, the *ids* of the database images containing that word are retrieved from the inverted file. For each image *id* **j** retrieved, its score is incremented using Equation 4.5, where $m_i(j)$ is the term frequency of word **i** in image **j**. After processing all the visual words in the query image, the scores are normalized in order to obtain the final ranking. This scoring system gives the dot products between the document vectors of the query and database images.

$$score(j) += m_i(j) \cdot w_i \tag{4.5}$$

4.2 Experimental Results

The searching scheme was tested by performing queries on a database of 118 paracatadioptric images (CATAset). Perspective images divided in sets A, B, C and D as described in Section 2.2, were used as query images (451 query images). The retrieval performance was evaluated for varying amounts of training data (for the visual vocabulary construction), and for different vocabulary sizes. The keypoint extraction in the query images is done using SIFT algorithm with the initial scale doubled (see Section 3.1).

Figure 4.3 shows the retrieval results obtained (extensive results are reported in Table 4.1). The global retrieval performance is measured through the percentage of cases where the correct database image is retrieved in first place (Figure 4.3(a)), in the TOP 3 (Figure 4.3(b)) and in the TOP 5 (Figure 4.3(c)) of the retrieved images. Two different data sets for building the visual vocabulary were used: PER10200 with 10200 perspective images (used in [25]) and MIX900 with 591 perspective images from PER10200 and 309 paracatadioptric images, different from the database images. The vocabulary sizes tested were: 10^5, 10^6 and 11^6. Also, for keypoint extraction in the database images, the SIFT (MIX900-SIFT and PER10200-SIFT) and *CylSIFT* (MIX900-CylSIFT and PER10200-CylSIFT) algorithms are compared.

From Figure 4.3 it can be oberserved that the most suitable vocabulary size is 10^6. Figure 4.3(a), in some cases shows a small improvement of the correct retrieval percentage when using 11^6 visual words relatively to 10^6. However, in addition to a greater computational effort, there is a reduction in the TOP 3 and TOP 5 retrieval for all the cases. In general, the retrieval performance increases with the number of visual words. Although, by using a vocabulary too large, the variability and noise in the descriptor vectors frequently move them between different quantization cells. There is a trade-off between distinctiveness (larger vocabulary) and repeatability (smaller vocabulary).

About the training data, it seems that by using mixtures of perspective and omnidirectional images we are able to create vocabularies from less images that are almost as efficient as vocabularies from a larger number of perspective images. However, to be sure about this, we should perform further experiments.

As it can be seen, the best retrieval percentage was 57.2% (PER10200-SIFT, 10^6

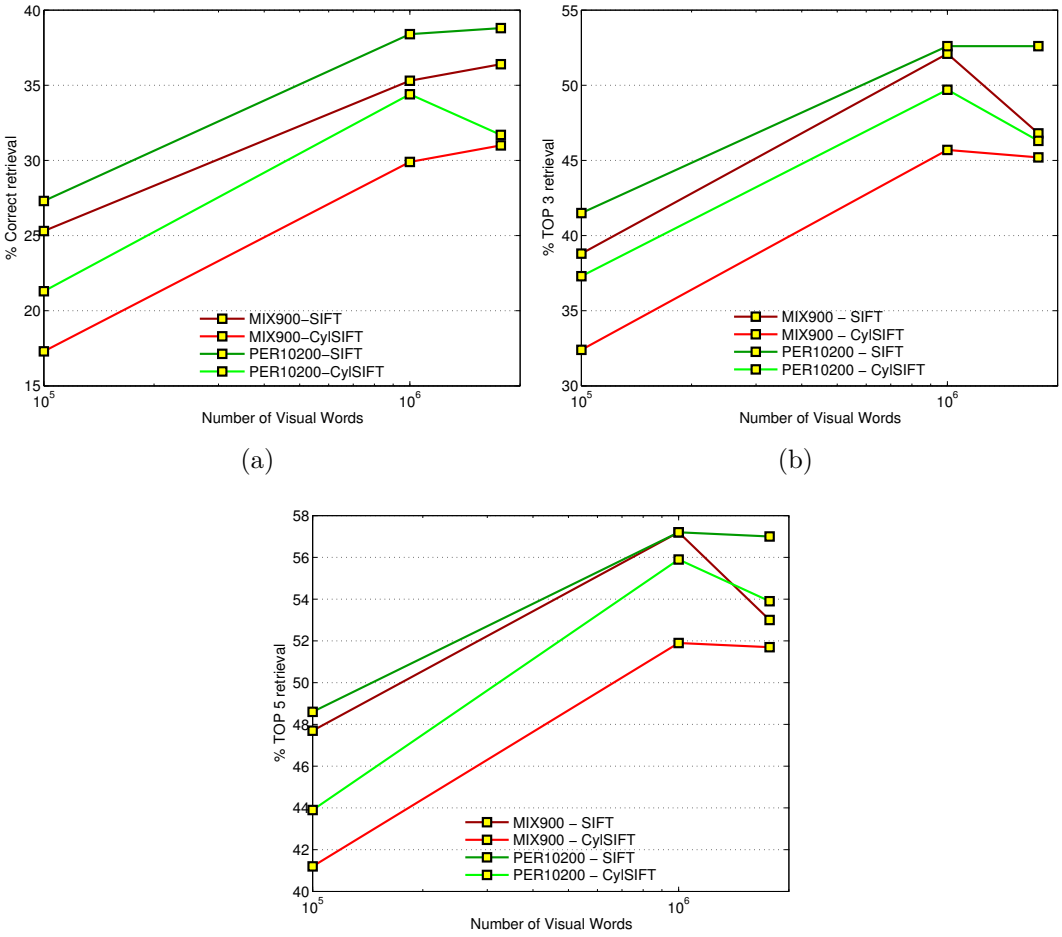


Figure 4.3: Retrieval results using different visual vocabularies and keypoints extraction algorithms.

visual words, TOP 5). Obviously, these retrieval percentages depend on the query images used. In the tests, 239 of the 451 query images are from sets C and D, which originate few matched visual words due to their high difficulty level, resulting in lower retrieval percentages.

To further conclusions, it is important to understand in which circumstances a correct match between two images can, or cannot, lead to a common visual word. In the retrieval tests performed, it was observed that the number of visual words in common between a query image and the corresponding database image is lower than the number of correct matches between these images. Therefore, there are correct matches that give raise to different visual words. This fact is due to the euclidean distance between the descriptor vectors. While in the previous chapters, two descriptors were considered as a correct

Retrieval Results				
	Voc. size	**1**	**3**	**5**
MIX900-SIFT	10^5	25.3	38.8	47.7
	10^6	35.3	52.1	57.2
	11^6	36.4	46.8	53.0
PER10200-SIFT	10^5	27.3	41.5	48.6
	10^6	**38.4**	**52.6**	**57.2**
	11^6	38.8	52.6	57.0
MIX900-CylSIFT	10^5	17.3	32.4	41.2
	10^6	29.9	45.7	51.9
	11^6	31.0	45.2	51.7
PER10200-CylSIFT	10^5	21.3	37.3	43.9
	10^6	**34.4**	**49.7**	**55.9**
	11^6	31.7	46.3	53.9

Table 4.1: Retrieval results using different visual vocabularies and keypoints extraction algorithms. The columns indicate: Voc. size - number of visual words; 1 - percentage of cases with correct retrieval; 3 - percentage of cases where the correct database image was retrieved in the TOP 3 of the retrieved images; 5 - percentage of cases where the correct database image was retrieved in the TOP 5 of the retrieved images.

match using the ratio between the distances to the first and second nearest neighbor, 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. Several cases were carefully studied and, as expected, in general the euclidean distance between matches corresponding to the same visual word is lower than the distance between matches with different visual words.

The results of Figure 4.3 show that the scoring methods using SIFT are slightly better than the ones using *CylSIFT*. To understand these results, the cases where the correct database image is retrieved in first place were analyzed when both methods retrieved correctly, and when one of them did not. Figure 4.4 shows the results obtained. As it can be seen, there is a relevant number of cases where the retrieval scheme using *CylSIFT* fails and the one using SIFT does not (second bar). Therefore, these cases were carefully analyzed and it was observed (see Table 4.2) that, the average number of common words between the query and the correct database image was higher when using SIFT, which explain the results obtained. However, the number of correct matches

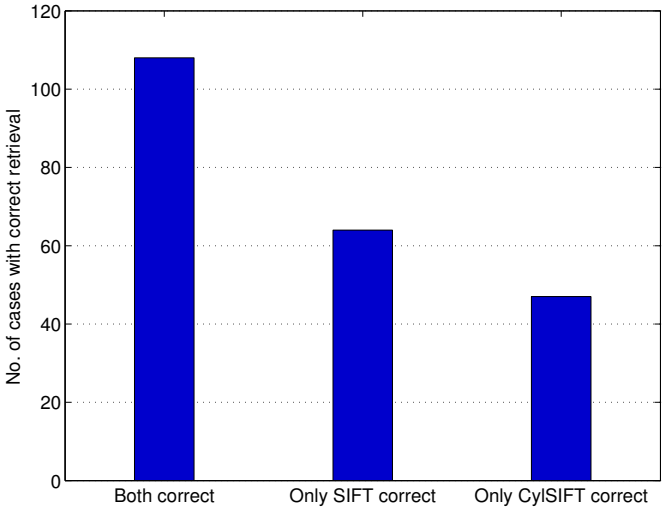


Figure 4.4: Number of cases where the correct database image is retrieved in first place by: both PER10200-SIFT and PER10200-CylSIFT; only PER10200-SIFT; and only PER10200-CylSIFT.

was higher when using *CylSIFT*, which was the expected due to the results obtained in Chapter 3. Therefore, for these cases, an higher number of correct matches did not led to an higher number of common visual words. This occurred because the average distances between the correct matches using SIFT were lower than the ones using *CylSIFT*.

Sift		*CylSIFT*		
Correct matches	**VW**	**Correct matches**	**VW-correct**	**VW-retrieved**
47.9	9.9	60.1	7.8	12.2

Table 4.2: Average number of correct matches and common visual words between the query and the correct database image, for the cases where the retrieval scheme using *CylSIFT* fails and the one using SIFT does not. For *CylSIFT*, the average number of common visual words between the query and the retrieved image, is also presented (VW-retrieved).

4.3 Improving Retrieval Performance: Geometry-Preserving Visual Phrases

In most state-of-the-art retrieval technologies, the database images are represented as histograms of visual words. As explained in the previous sections, in the searching step similar images are retrieved from the database and a ranking is generated. To establish

this ranking, an inverted file structure is used in order to facilitate fast access to images with common words. Unfortunately, this model does not take into account spatial information. In [33], Yimeng Zhang *et al.* propose to encode spatial information in the searching step by using geometry-preserving visual phrases (GVP). A GVP is constituted by a group of visual words in a particular spatial layout. This method can provide an initial ranking with more spatial information.

The objective of using GVP is to take into account the spatial relations between visual words. As referred in [33], a set of k visual words in a certain spatial layout define a GVP of length k. Figure 4.5 illustrates how the co-occurring GVP in two images are identified. For each pair of the same word in the images, the offset is computed by subtracting their corresponding locations. As it can be seen, 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. k votes in the same offset cell correspond to a co-occurring GVP of length k. In Figure 4.5, words A,B and C correspond to a co-occurring GVP of length 3.

Therefore, rather than keep one entry for each image to accumulate the scores, M cells are kept for each image, where M is the number of possible offsets. The voting procedure to obtain the similarity scores of length k-GVP is described as follows:

1. M cells for each database image are initialized to 0. Each cell represents an offset value.

2. For each word **j** in the query image, the *ids* and locations of the occurrences of **j** in the database images are retrieved through the inverted file. For each retrieved word occurrence **d** in image **i**, the offset between **j** and **d** is computed and the corresponding offset cell of image **i** is incremented [33]:

$$S_{i,x_d-x_j,y_d-y_j} += 1 \tag{4.6}$$

where (x_d, y_d) and (x_j, y_j) are the locations, in the offset space, of the words **d** and **j**, respectively. S_i are the scores of the database image **i**. As mentioned in [33], in addition to the number of words $S_{i,m}$ in each offset cell m, it is necessary to keep the sum of the weights of these words $D_{i,m}$:

$$D_{i,x_d-x_j,y_d-y_j} \mathrel{+}= w_j \tag{4.7}$$

3. The final scores for the database images are computed, considering just the offset cells m with value equal or greater than k (number of co-occurring GVP of length k, with weighting):

$$\hat{S}_i = \sum_{m>=k} D_{i,m} \binom{S_{i,m}-1}{k-1} \tag{4.8}$$

4. Finally, the scores for the database images are normalized:

$$\hat{S}_i^* = \frac{\hat{S}_i}{\sum_i \hat{S}_i} \tag{4.9}$$

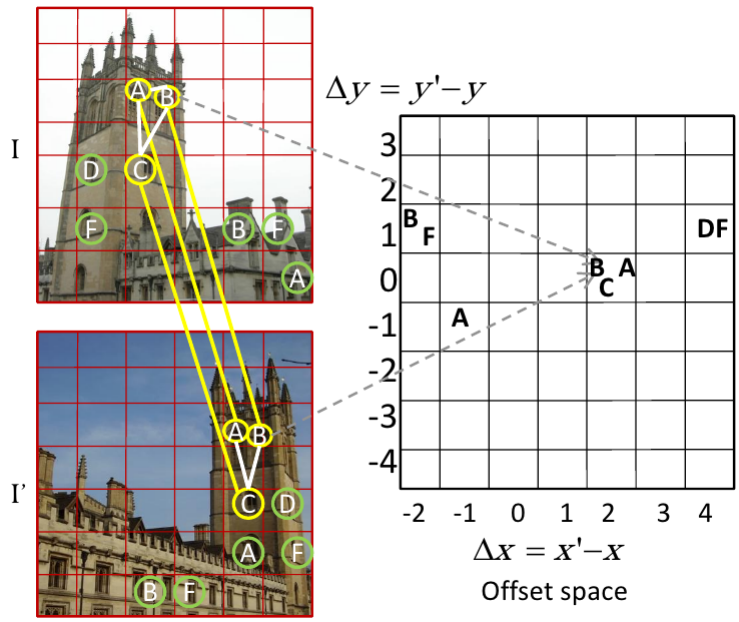


Figure 4.5: Identification of the co-occurring GVP in a pair of images [33].

It is important to refer that in [33], Yimeng Zhang *et al.* work with perspective images only. In order to maintain coherence between the coordinates of the query and the database images, the locations of the words in the paracatadioptric images are converted to polar coordinates (see Equation 3.1). This is done before the computation of the spatial offsets. It was observed experimentally that using this conversion improves the retrieval

results.

4.3.1 Experimental results

This approach is evaluated using the retrieval schemes PER10200-SIFT and PER10200-CylSIFT with 10^6 visual words. As in [33], the image space is quantized by a 10x10 matrix constituting the offset space. Using an higher dimension leads to a more rigorous spatial modeling, although, more memory and computational effort would be necessary. For the voting procedure, GVP of length 2 ($k = 2$) are used. As showed in [33], length 2 is the best among the lengths from 1 to 5.

The results of Table 4.3 show an improvement in the retrieval scheme using *CylSIFT* relatively to the results obtained in Section 4.2. This means that the words in the database are spatially consistent when compared with SIFT (SIFT had improvements only in the correct retrieval percentage). It can be concluded that by using the spatial information of the words when using *CylSIFT*, the results are improved. Analyzing Table 4.3, the best retrieval results are obtained when using *CylSIFT* with GVP.

	BoV			**GVP**			**Variation**		
	1	**3**	**5**	**1**	**3**	**5**	**1**	**3**	**5**
SIFT	38.4	52.6	57.2	41.0	51.9	55.7	+2.6	-0.7	-1.5
CylSIFT	34.4	49.7	55.9	41.5	57.4	61.4	+7.1	+7.7	+5.5

Table 4.3: Retrieval results using **GVP** in the retrieval schemes PER10200-SIFT and PER10200-CylSIFT with 10^6 visual words. **BoV** corresponds to the results obtained in Section 4.2. The columns indicate: 1 - percentage of cases with correct retrieval; 3 - percentage of cases where the correct database image was retrieved in the TOP 3 of the retrieved images; 5 - percentage of cases where the correct database image was retrieved in the TOP 5 of the retrieved images.

4.4 Geometric Consistency Check

In order to provide a more precise ranking of the retrieved images, a re-ranking of the top-retrieved images can be made [27]. This *post-processing* step, consists on a geometric consistency check performed with RANSAC. Using the extracted keypoint descriptors, matches between the query and the top-retrieved images are established. Then, RANSAC

is used to robustly fit an homography between the views. The re-ranking is done based on the number of correct matches obtained. This final geometric consistency check was performed over the top 5 and top 10 retrieved images. Table 4.4 and Table 4.5 show the results obtained.

Re-ranking Results			
No. images	**1**	**3**	**5**
5	58.5	61.4	61.4
10	63.2	66.7	67.6

Table 4.4: Retrieval results after the final consistency check in the best retrieval scheme. From left to right, the columns indicate: No. images - Number of top-ranked images used for re-ranking; 1 - percentage of cases with correct retrieval; 3 - percentage of cases where the correct database image was retrieved in the TOP 3 of the retrieved images; 5 - percentage of cases where the correct database image was retrieved in the TOP 5 of the retrieved images.

Retrieval % - Best Searching Scheme			
	1	**3**	**5**
Set A	72.6	76.4	76.4
Set B	82.1	84.0	84.0
Set C	61.3	62.3	63.2
Set D	42.1	48.9	51.1

Table 4.5: Retrieval results after the final consistency check for the top 10 retrieved images, by the best retrieval scheme. This table shows the individual retrieval percentage for the different perspective images sets. From left to right, the columns indicate: 1 - percentage of cases with correct retrieval; 3 - percentage of cases where the correct database image was retrieved in the TOP 3 of the retrieved images; 5 - percentage of cases where the correct database image was retrieved in the TOP 5 of the retrieved images.

It was observed that using the geometric consistency check leads to a boost of performance. To resume, in this scheme, the visual vocabulary is built by feature extraction using SIFT algorithm, over 10200 perspective images (data set PER10200). The vocabulary is based on a hierarchical tree structure and is constituted by 10^6 visual words. For feature extraction on the database images, *CylSIFT* algorithm is used. The scoring system is based on geometry-preserving visual phrases. Figure 4.6 shows the retrieval performance of this scheme, comparing the typical *bag of visual words* scoring system

(Section 4.1.3), GVP and also GVP with re-ranking of the top 5 and top 10 retrieved images. Figure 4.7 shows examples of cases where there is retrieval of the correct database image.

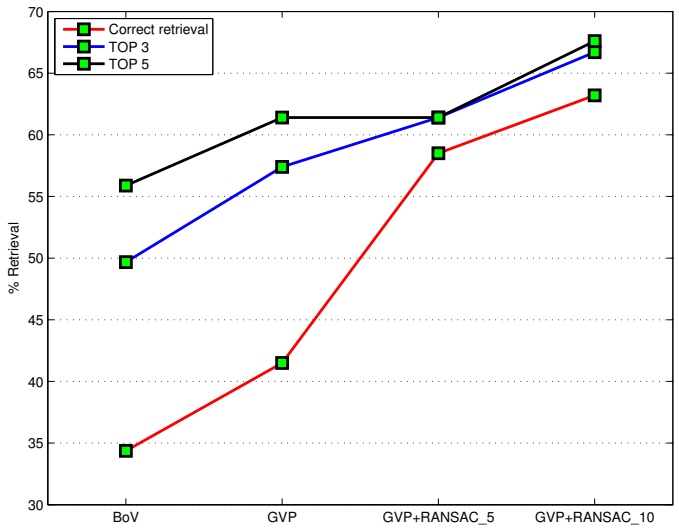


Figure 4.6: Retrieval performance of the best searching scheme. The red, blue and black curves show the percentage of cases with correct retrieval, percentage of cases where the correct database image was retrieved in the TOP 3, and in the TOP 5 of the retrieved images, respectively. Different methods are compared: the typical *bag of visual words* scoring system (Section 4.1.3), GVP and also GVP with re-ranking of the top 5 and top 10 retrieved images.

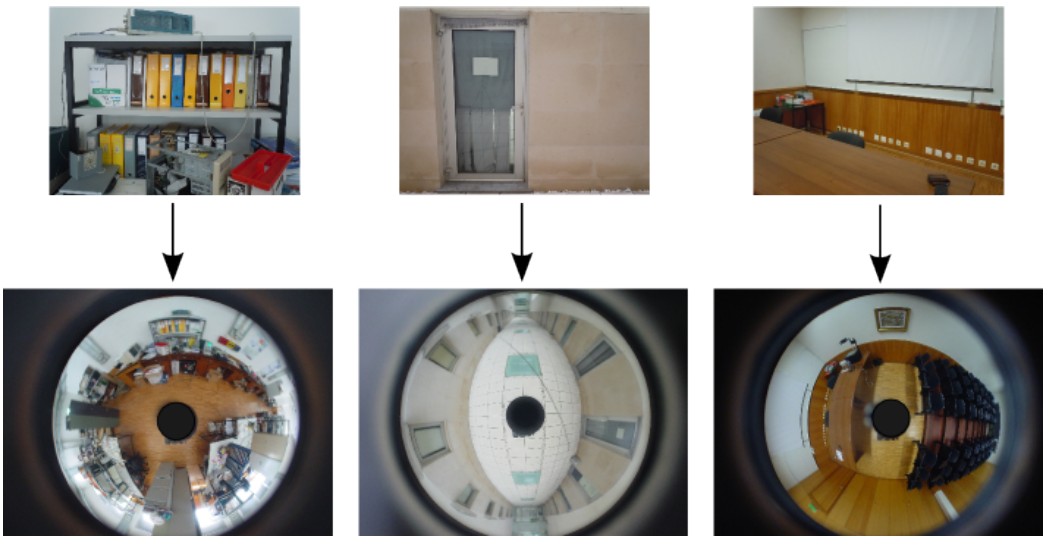


Figure 4.7: Examples of queries where there is retrieval of the correct database image.

The main factors that affect the retrieval performance are:

- Low number of matches between the query and the corresponding database image (see Figure 4.8(a)).
- Perceptual aliasing. Similar structures/scenes in the environment affect the recognition scheme by confusing it (see Figure 4.8(b)).
- Scale invariant feature detection and description that cannot cope with strong viewpoint and/or scale changes (see Figure 4.8(c)).

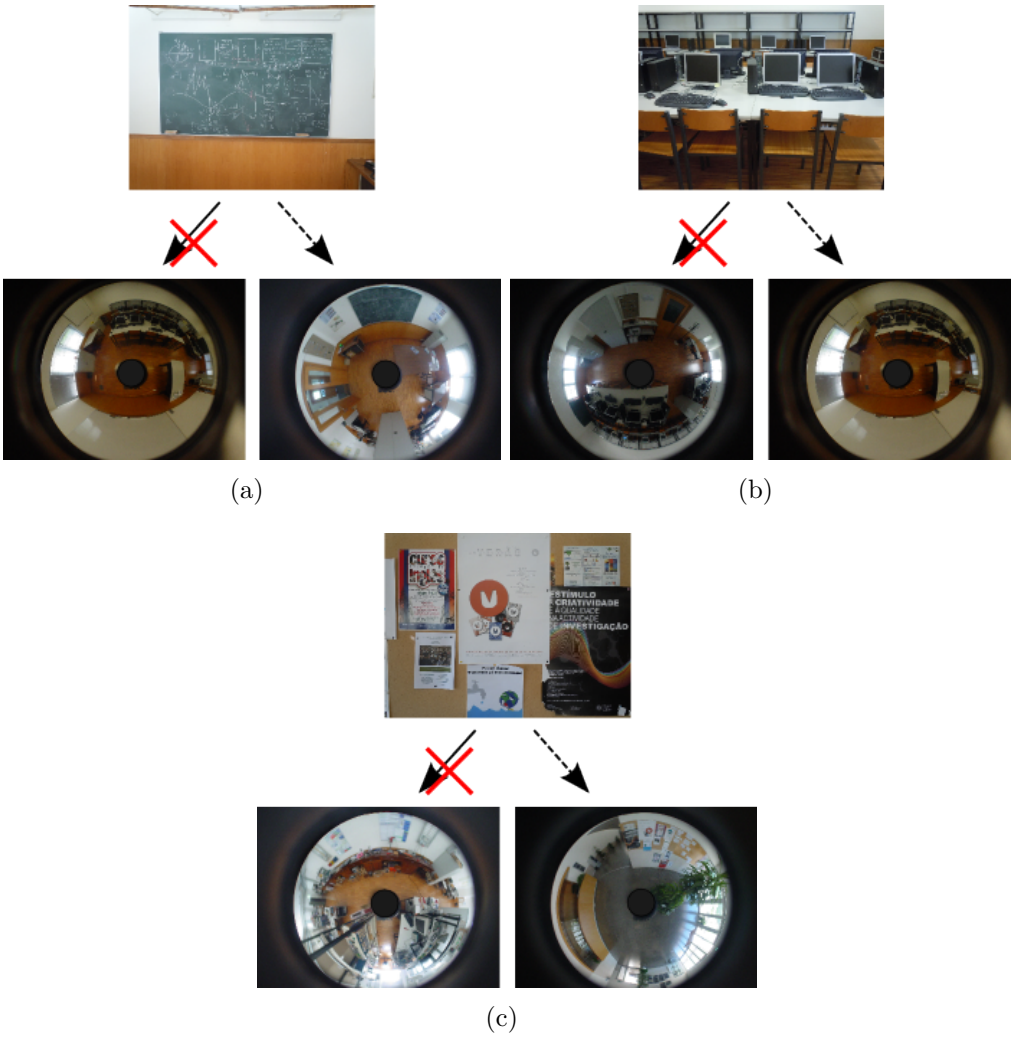


Figure 4.8: Examples of queries where there is retrieval of the incorrect database image. The main reasons for incorrect retrieval are: low number of matches (a); perceptual aliasing (b); and strong viewpoint/scale changes (c).

Additionally, Table 4.5 shows that the percentage of correct retrieval for images from sets C and D are much lower than with sets A and B. As referred before, sets C and D,

in general, originate few matched visual words due to their high difficulty level, resulting in lower retrieval percentages.

Chapter 5

Conclusions and Outlook

This thesis conducts an exploratory work with the objective of developing techniques for visual recognition in panoramic images using perspectives for the query. The main challenge was the establishment of correspondences between a query image taken with a standard camera (described by the pin-hole model) and a database of paracatadioptric images.

Two images corresponding to the same scene/object can be matched by establishing correspondences between their points of interest, also called keypoints. SIFT algorithm, originally proposed by David Lowe, is one of the most robust approaches for keypoint extraction in terms of scale, rotation and viewpoint invariance. However, SIFT takes into account neither the geometry, nor the radial distortion of paracatadioptric images, penalizing the performance of the image analysis applications that use these images. Therefore, different types of rectification strategies were studied in order to correct the radial distortion present in the paracatadioptric images. Matching between perspective images and rectified images from the paracatadioptric images was performed. From the results obtained it was verified that the best rectification method was to transform the paracatadioptric images to cylindrical coordinates.

In order to improve the results of matching between perspective and paracatadioptric images, several modifications to the original SIFT algorithm were proposed:

- A change in the initial scale for the Gaussian blurring (from 1.6 to 3.2). The objective is to eliminate higher frequency features in the perspective images, which generate many false matches.

- Construction of the scale-space using adaptive filtering that takes into account the geometry of the paracatadioptric images and compensates the radial distortion (the values for the Gaussian filter are computed using a mapping to cylindrical coordinates - *CylSIFT*). This approach outperformed the competing methods showing the best results in terms of average number and percentage of correct matches.

As an additional result, a brief study of feature detection and matching between paracatadioptric images was performed. SIFT algorithm was used, along with different rectification methods. Also, a new mapping function for the adaptive filtering (mapping to a rectified perspective *view*) was tested, outperforming the other approaches.

To efficiently search the database of paracatadioptric images, a recognition scheme using an hierarchical vocabulary tree was built, based on [25]. Different visual vocabularies and training data were used. Additionally, several methods for feature extraction in the database images were analyzed, including the original SIFT algorithm and SIFT with implicit filtering. The results obtained showed that *CylSIFT* did not outperform original SIFT. There are two main reasons that can explain these results:

- The quantization of the descriptors vectors is based on the Euclidean distance to the clusters centers of the vocabulary tree. In Chapter 3, a pair of descriptors was considered a correct match based on the ratio between the distances to the first and second nearest neighbor, which is a different approach. Therefore, a correct match may not correspond to common cluster centers.

- When using a vocabulary tree, the spatial information is discarded, which is prejudicing *CylSIFT* because the words obtained with this method are more spatially consistent than with SIFT.

Finally, methods to improve the retrieval performance were studied. To encode more spatial information in the searching step, the new concept of geometry-preserving visual phrases was used [33]. The results showed an improvement in the retrieval scheme using *CylSIFT* as feature extraction algorithm for the database images. This means that the *CylSIFT* words are spatially more consistent than those of SIFT. Additionally, to provide a more precise ranking of the retrieved images, a geometric consistency check (using

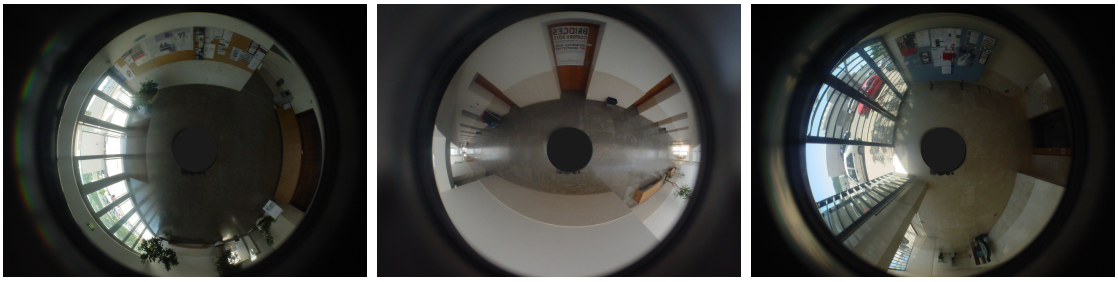
RANSAC) was performed on the top-ranked images. This *post-processing* step gives a significant boost in the retrieval performance.

Appendix A

Images used in the tests

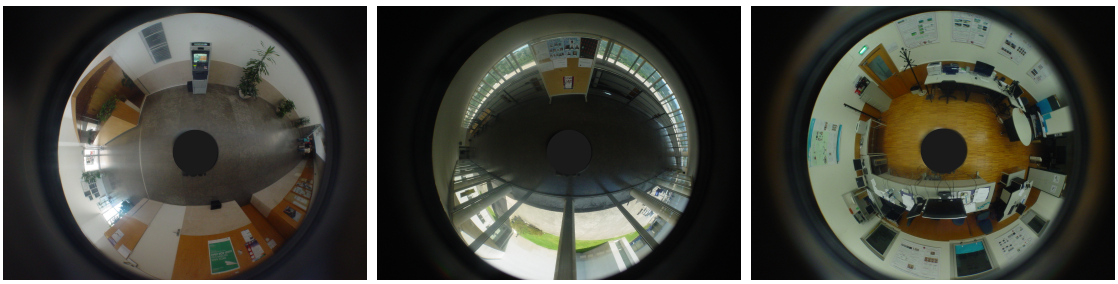
This appendix shows all the images used in the tests of Chapters 2 and 3.

(a) cata1 (b) cata2 (c) cata3



(d) cata4 (e) cata5 (f) cata6

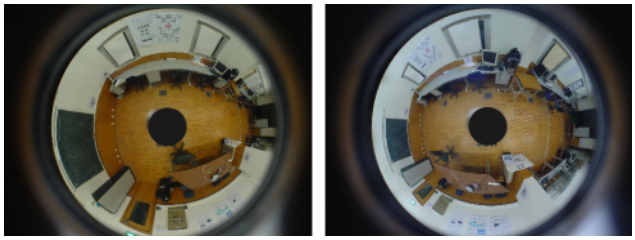
(g) cata7 (h) cata8 (i) cata9



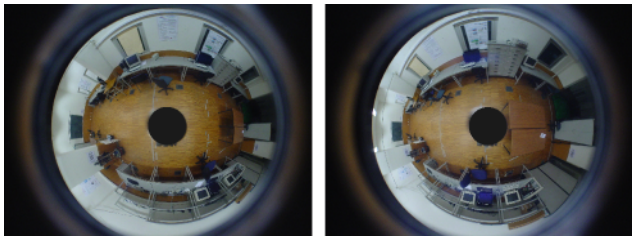
(j) cata10 (k) cata11 (l) cata12

(m) cata13

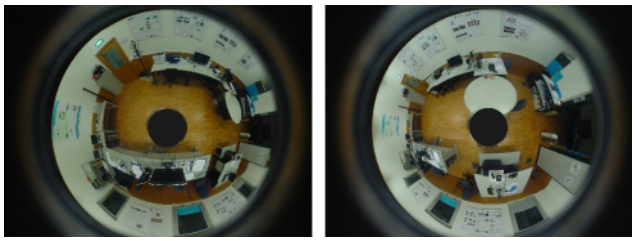
Figure A.1: Set of paracatadioptric images used for feature detection and matching in Sections 2.4 and 3.4.



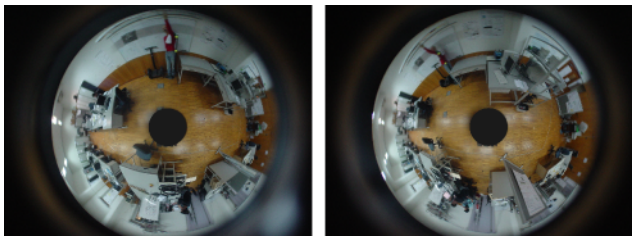
(a) Par 1



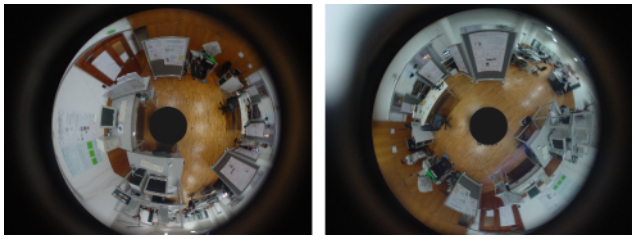
(b) Par 2



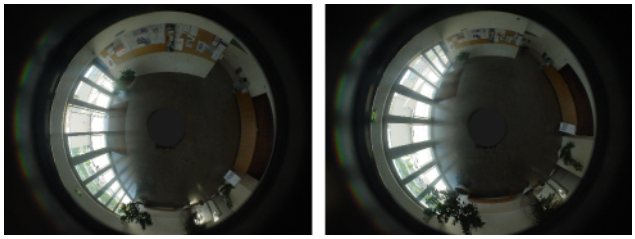
(c) Par 3



(d) Par 4



(e) Par 5



(f) Par 6

Figure A.2: Pairs of paracatadioptric images used for feature detection and matching in Section 3.5.

Appendix B

Full Tables of Results

This appendix shows the full tables with the results of feature detection and matching (Chapters 2 and 3).

		Inliers/Matches			
		Sift	**VCP with offset**	**VCP aligned**	**Cylinder**
cata1	**Set A**	**128/230 (55.65%)**	49/168 (29.17%)	110/229 (48.03%)	120/270 (44.44%)
	Set B	**213/455 (46.81%)**	45/396 (11.36%)	199/429 (46.39%)	198/478 (41.42%)
	Set C	**163/464 (35.13%)**	31/384 (8.07%)	140/417 (33.57%)	144/431 (33.41%)
	Set D	9/68 (13.24%)	2/53 (3.77%)	8/64 (12.50%)	**15/63 (23.81%)**
cata2	**Set A**	107/359 (29.81%)	90/303 (29.70%)	144/349 (41.26%)	**146/329 (44.38%)**
	Set B	95/255 (37.25%)	44/212 (20.75%)	103/254 (40.55%)	**130/266 (48.87%)**
	Set C	4/492 (0.81%)	0/508 (0%)	**11/498 (2.21%)**	3/469 (0.64%)
	Set D	78/308 (25.32%)	64/299 (21.40%)	81/319 (25.39%)	**82/298 (27.52%)**
cata3	**Set A**	359/899 (39.93%)	182/816 (22.30%)	303/804 (37.69%)	**491/1053 (46.63%)**
	Set B	500/1264 (39.56%)	244/1061 (23.00%)	395/1047 (37.73%)	**583/1382 (42.19%)**
	Set C	332/1152 (28.82%)	136/1094 (12.43%)	241/987 (24.42%)	**408/1247 (32.72%)**
	Set D	139/656 (21.19%)	102/614 (16.61%)	125/598 (20.90%)	**232/791 (29.33%)**
cata4	**Set A**	85/230 (36.96%)	18/196 (9.18%)	83/248 (33.47%)	**95/261 (36.40%)**
	Set B	89/257 (34.63%)	20/212 (9.43%)	95/267 (35.58%)	**115/279 (41.22%)**
	Set C	4/314 (1.27%)	0/339 (0%)	35/350 **(10%)**	**38/1290** (2.95%)
	Set D	14/186 (7.53%)	13/187 (6.95%)	18/203 (8.87%)	**20/166 (12.05%)**
cata5	**Set A**	26/271 (9.59%)	25/220 (11.36%)	2/327 (0.61%)	**42/266 (15.79%)**
	Set B	94/377 (24.93%)	40/278 (14.39%)	74/431 (17.17%)	**105/415 (25.30%)**
	Set C	**65/604 (10.76%)**	24/484 (4.96%)	41/733 (5.59%)	49/664 (7.38%)
	Set D	82/300 (27.33%)	30/232 (12.93%)	61/358 (17.04%)	**91/322 (28.26%)**
cata6	**Set A**	19/164 (11.59%)	15/171 (8.77%)	**24/162 (14.81%)**	21/169 (12.43%)
	Set B	**88/322 (27.33%)**	57/305 (18.69%)	76/307 (24.76%)	78/298 (26.17%)
	Set C	**66/395 (16.71%)**	44/381 (11.55%)	55/413 (13.32%)	36/373 (9.65%)
	Set D	46/254 (18.11%)	26/214 (12.15%)	**52/249 (20.88%)**	49/238 (20.59%)
cata7	**Set A**	11/117 (9.40%)	11/122 (9.02%)	**24/129 (18.60%)**	16/125 (12.80%)
	Set B	61/204 (29.90%)	55/221 (24.89%)	67/261 (25.67%)	**73/231 (31.60%)**
	Set C	19/409 (4.65%)	14/446 (3.14%)	**24/489 (4.91%)**	12/424 (2.83%)
	Set D	11/84 (13.10%)	14/95 (14.74%)	**25/102 (24.51%)**	12/101 (11.88%)
cata8	**Set A**	113/393 (28.75%)	9/402 (2.24%)	110/487 (22.59%)	**128/429 (29.84%)**
	Set B	**116/410 (28.29%)**	23/383 (6.01%)	115/462 (24.89%)	109/407 (26.78%)
	Set C	**2/463 (0.43%)**	2/581 (0.34%)	0/607 (0%)	0/527 (0%)
	Set D	55/395 (13.92%)	48/468 (10.26%)	58/434 (13.36%)	**100/456 (21.93%)**
cata9	**Set A**	**59/198 (29.80%)**	40/187 (21.39%)	48/199 (24.12%)	58/187 (31.02%)
	Set B	54/201 (26.87%)	35/218 (16.06%)	57/231 (24.68%)	**63/199 (31.66%)**
	Set C	0/399 (0%)	0/379 (0%)	0/355 (0%)	**10/368 (2.72%)**
	Set D	48/137 (35.04%)	31/133 (23.31%)	37/146 (25.34%)	**48/132 (36.36%)**
cata10	**Set A**	56/169 (33.14%)	52/156 (33.33%)	55/234 (23.50%)	**69/199 (34.67%)**
	Set B	61/237 (25.74%)	50/211 (23.70%)	53/266 (19.92%)	**71/262 (27.10%)**
	Set C	4/215 (1.86%)	24/174 **(13.79%)**	2/255 (0.78%)	**26/236** (11.02%)
	Set D	34/251 (13.55%)	12/309 (3.88%)	**44/253 (17.39%)**	32/259 (12.36%)
cata11	**Set A**	88/292 (30.14%)	68/275 (24.73%)	86/289 **(29.76%)**	**109/401** (27.18%)
	Set B	127/408 (31.13%)	83/363 (22.87%)	126/405 (31.11%)	**137/401 (34.16%)**
	Set C	83/448 (18.53%)	73/453 (16.11%)	86/450 (19.11%)	**126/453 (27.81%)**
	Set D	15/224 (6.70%)	0/262 (0.00%)	19/241 (7.88%)	**33/254 (12.99%)**
cata12	**Set A**	100/256 (39.06%)	62/242 (25.62%)	101/264 (38.26%)	**107/267 (40.07%)**
	Set B	104/324 (32.10%)	33/308 (10.71%)	111/331 (33.53%)	**117/315 (37.14%)**
	Set C	**21/335 (6.27%)**	0/331 (0%)	20/323 (6.19%)	11/295 (3.73%)
	Set D	15/213 (7.04%)	2/218 (0.92%)	44/212 (20.75%)	**26/217 (11.98%)**
cata13	**Set A**	114/390 (29.23%)	68/330 (20.61%)	129/428 (30.14%)	**141/387 (36.43%)**
	Set B	111/353 (31.44%)	75/320 (23.44%)	115/418 (27.51%)	**123/362 (33.98%)**
	Set C	51/334 (15.27%)	29/338 (8.58%)	24/349 (6.88%)	**61/370 (16.49%)**
	Set D	16/315 (5.08%)	3/338 (0.89%)	**25/370 (6.76%)**	16/313 (5.11%)

Table B.1: Results of feature matching between the rectified views (virtual camera planes (VCP) with offset, VCP aligned and cylindrical images) from a set of 13 paracatadioptric images and the corresponding perspective images, using SIFT algorithm. Matching results using the original SIFT algorithm on the reflected paracatadioptric images is also presented (*Sift*). For each case, the number of inliers (correct matches), total number of matches and inliers percentage is represented. The values in bold correspond to the best results for each case.

	Detections (rectified images)			
	Sift	**VCP with offset**	**VCP aligned**	**Cylinder**
cata1	637	370	420	809
cata2	2188	1344	817	3635
cata3	2649	779	939	3280
cata4	1266	303	249	1290
cata5	469	230	101	422
cata6	1524	800	563	2176
cata7	770	128	96	777
cata8	1525	218	390	1801
cata9	1663	519	115	1841
cata10	887	300	272	920
cata11	1611	369	385	1900
cata12	1556	394	425	2283
cata13	1608	332	432	2504

Table B.2: Results of feature matching between the rectified views (VCP with offset, VCP aligned and cylindrical images) from a set of 13 paracatadioptric images and the corresponding perspective images, using SIFT algorithm. Matching results using the original SIFT algorithm on the reflected paracatadioptric images is also presented (*Sift*). For each case, the number of detections is represented.

		Inliers/Matches			
		Sift-LF	*Polar*	*Cylinder-LF*	*CylSIFT*
cata1	**Set A**	136/219 (62.10%)	92/168 (54.76%)	131/234 (55.98%)	**141/203 (69.46%)**
	Set B	210/349 (60.17%)	156/289 (53.98%)	183/312 (58.65%)	**235/319 (73.67%)**
	Set C	182/346 (52.60%)	140/323 (43.34%)	164/359 (45.68%)	**195/308 (63.31%)**
	Set D	14/70 (20.00%)	14/54 (25.93%)	**16/55 (29.09%)**	14/50 (28.00%)
cata2	**Set A**	139/270 (51.48%)	143/247 (57.89%)	**185/269** (68.77%)	154/203 **(75.86%)**
	Set B	129/193 (66.84%)	135/196 (68.88%)	**173/231** (74.89%)	147/196 **(75.00%)**
	Set C	31/106 (29.25%)	40/108 (37.04%)	40/91 (43.96%)	**41/91 (45.05%)**
	Set D	66/187 (35.29%)	80/195 (41.03%)	85/174 (48.85%)	**85/136 (62.50%)**
cata3	**Set A**	320/597 (53.60%)	363/611 (59.41%)	432/692 (62.43%)	**435/650 (66.92%)**
	Set B	499/855 (58.36%)	509/845 (60.24%)	573/914 (62.69%)	**600/949 (63.22%)**
	Set C	351/763 (46.00%)	354/697 (50.79%)	458/838 (54.65%)	**508/811 (62.64%)**
	Set D	156/423 (36.88%)	166/406 (40.89%)	238/513 (46.39%)	**242/392 (61.73%)**
cata4	**Set A**	66/164 (40.24%)	90/146 (61.64%)	104/172 (60.47%)	**149/205 (72.68%)**
	Set B	103/173 (59.54%)	105/156 (67.31%)	119/180 (66.11%)	**167/220 (75.91%)**
	Set C	43/136 (31.62%)	43/128 (33.59%)	52/145 (35.86%)	**58/131 (44.27%)**
	Set D	16/102 (15.69%)	25/423 (5.91%)	25/423 (5.91%)	**28/63 (44.44%)**
cata5	**Set A**	47/125 (37.60%)	40/118 (33.90%)	52/130 (40.00%)	**100/156 (64.10%)**
	Set B	89/193 (46.11%)	90/191 (47.12%)	99/204 (48.53%)	**183/243 (75.31%)**
	Set C	65/393 (16.54%)	83/401 (20.70%)	98/437 (22.43%)	**169/342 (49.42%)**
	Set D	77/133 (57.89%)	77/140 (55.00%)	85/141 (60.28%)	**129/164 (78.66%)**
cata6	**Set A**	41/106 (38.68%)	41/96 (42.71%)	40/101 (39.60%)	**51/91 (56.04%)**
	Set B	89/196 (45.41%)	89/184 (48.37%)	89/186 (47.85%)	**161/204 (78.92%)**
	Set C	77/219 (35.16%)	74/196 (37.76%)	62/192 (32.29%)	**124/182 (68.13%)**
	Set D	39/146 (26.71%)	42/140 (30.00%)	42/127 (33.07%)	**74/123 (60.16%)**
cata7	**Set A**	20/61 (32.79%)	27/68 (39.71%)	**31/70 (44.29%)**	22/61 (36.07%)
	Set B	65/127 (51.18%)	70/134 (52.24%)	87/145 (60.00%)	**92/148 (62.16%)**
	Set C	51/193 (26.42%)	56/224 (25.00%)	60/219 (27.40%)	**60/164 (36.59%)**
	Set D	11/56 (19.64%)	15/59 (25.42%)	16/67 (23.88%)	**17/56 (30.36%)**
cata8	**Set A**	120/241 (49.79%)	127/241 (52.70%)	139/271 (51.29%)	**140/228 (61.40%)**
	Set B	115/235 (48.94%)	112/249 (44.98%)	130/237 (54.85%)	**151/211 (71.56%)**
	Set C	47/220 (21.36%)	47/209 (22.49%)	52/241 (21.58%)	**52/120 (43.33%)**
	Set D	46/186 (24.73%)	42/199 (21.11%)	65/207 (31.40%)	**69/138 (50.00%)**
cata9	**Set A**	66/111 (59.46%)	53/110 (48.18%)	68/112 (60.71%)	**96/145 (66.21%)**
	Set B	62/134 (46.27%)	55/131 (41.98%)	73/131 (55.73%)	**117/165 (70.91%)**
	Set C	46/145 (31.72%)	41/149 (27.52%)	51/131 (38.93%)	**79/152 (51.97%)**
	Set D	28/67 (41.79%)	28/72 (38.89%)	**33/59 (55.93%)**	33/65 (50.77%)
cata10	**Set A**	73/140 (52.14%)	74/147 (50.34%)	83/136 (61.03%)	**101/152 (66.45%)**
	Set B	71/152 (46.71%)	71/162 (43.83%)	87/149 (58.39%)	**113/177 (63.84%)**
	Set C	51/135 (37.78%)	51/144 (35.42%)	68/159 (42.77%)	**73/159 (45.91%)**
	Set D	21/86 (24.42%)	26/81 (32.10%)	27/76 (35.53%)	**40/87 (45.98%)**
cata11	**Set A**	92/185 (49.73%)	99/188 (52.66%)	107/198 (54.04%)	**110/158 (69.62%)**
	Set B	129/242 (53.31%)	135/265 (50.94%)	151/261 (57.85%)	**188/263 (71.48%)**
	Set C	108/270 (40.00%)	116/268 (43.28%)	141/281 (50.18%)	**157/255 (61.57%)**
	Set D	12/153 (7.84%)	32/167 (19.16%)	**33/152** (21.71%)	32/101 **(31.68%)**
cata12	**Set A**	123/177 (69.49%)	128/203 (63.05%)	**150/211** (71.09%)	126/176 **(71.59%)**
	Set B	123/210 (58.57%)	124/210 (59.05%)	**140/220** (63.64%)	135/205 **(65.85%)**
	Set C	38/98 (38.78%)	50/105 (47.62%)	44/103 (42.72%)	**51/101 (50.50%)**
	Set D	38/115 (33.04%)	52/138 (37.68%)	43/115 (37.39%)	**53/103 (51.46%)**
cata13	**Set A**	128/237 (54.01%)	134/223 (60.09%)	147/229 (64.19%)	**171/221 (77.38%)**
	Set B	149/234 (63.68%)	150/219 (68.49%)	151/242 (62.40%)	**163/235 (69.36%)**
	Set C	75/199 (37.69%)	75/194 (38.66%)	78/184 (42.39%)	**83/169 (49.11%)**
	Set D	26/170 (15.29%)	22/170 (12.94%)	26/156 (16.67%)	**38/122 (31.15%)**

Table B.3: Results of feature matching between the perspective and the paracatadioptric/rectified images using different approaches (*Sift-LF*, *Polar*, *Cylinder-LF* and *CylSIFT*). For each case, the number of inliers (correct matches), total number of matches and inliers percentage is represented. The values in bold correspond to the best results for each case. The initial scale used on the construction of the scale-space for SIFT algorithm was doubled (1.6 to 3.2) for the perspective images.

	Detections (rectified image)			
	Sift-LF	*Polar*	*Cylinder-LF*	*CylSIFT*
cata1	632	642	809	793
cata2	2194	2441	3635	3089
cata3	2661	2592	3280	3809
cata4	1256	1218	1290	1722
cata5	475	424	422	659
cata6	1495	1534	2176	2100
cata7	773	738	777	897
cata8	1517	1537	1801	1860
cata9	1663	1676	1841	2182
cata10	884	867	920	1255
cata11	1637	1574	1900	2225
cata12	1550	1683	2283	2112
cata13	1612	1746	2504	2069

Table B.4: Results of feature matching between the perspective and the paracatadioptric/rectified images using different approaches (*Sift-LF*, *Polar*, *Cylinder-LF* and *CylSIFT*). For each case, the number of detections is represented. The initial scale used on the construction of the scale-space for SIFT algorithm was doubled (1.6 to 3.2) for the perspective images.

Bibliography

[1] Catpack toolbox, November 2008.

[2] J. J. Guerrero A. C. Murillo and C. Sagues. "surf features for efficient robot localization with omnidirectional images", 2007.

[3] J. Kosecka A. C. Murillo, P. Campos and J. J. Guerrero. Gist vocabularies in omnidirectional images for appearance based mapping and localization, 2010.

[4] Zafer Arican and Pascal Frossard. Scale Invariant Features and Polar Descriptors in Omnidirectional Imaging. *IEEE Transactions on Image Processing*, 2010.

[5] João P. Barreto. *General Central Projection Systems - Modeling, Calibration and Visual Servoing*. PhD thesis, University of Coimbra, Coimbra, September 2003.

[6] João P. Barreto. A unifying geometric representation for central projection systems. *Comput. Vis. Image Underst.*, 103:208–217, September 2006.

[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008.

[8] Mark Cummins and Paul Newman. FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[9] K. Daniilidis, A. Makadia, and T. Bulow. Image processing in catadioptric planes: spatiotemporal derivatives and optical flow computation. In *Omnidirectional Vision, 2002. Proceedings. Third Workshop on*, pages 3 – 10, 2002.

[10] Toon Goedemé, Tinne Tuytelaars, and Luc Van Gool. Omnidirectional sparse visual path following with occlusion-robust feature tracking. In *6th Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras, OMNIVIS05, in Conjunction with ICCV 2005*, 2005.

[11] P. Hansen, P. Corke, W. Boles, and K. Daniilidis. Scale-invariant features on the sphere. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1 –8, oct. 2007.

[12] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[13] Herve Jegou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Computer Vision and Pattern Recognition*, pages 1169–1176, 2009.

[14] P. D. Kovesi. MATLAB and Octave functions for computer vision and image processing. School of Computer Science & Software Engineering, The University of Western Australia. Available from: <http://www.csse.uwa.edu.au/∼pk/research/matlabfns/>.

[15] J. J. Guerrero L. Puig and P. Sturm. Matching of omnidirectional and perspective images using the hybrid fundamental matrix, 2008.

[16] Ouiddad Labbani-Igbida, Cyril Charron, and El Mustapha Mouaddib. Haar invariant signatures and spatial recognition using omnidirectional visual information only. *Auton. Robots*, 30:333–349, April 2011.

[17] Miguel Lourenco, Joao Pedro Barreto, and Francisco Vasconcelos. sRD-SIFT: Keypoint Detection and Matching in Images with Radial Distortion. *submitted to IEEE Transactions on Robotics*, to appear.

[18] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[19] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In *In British Machine Vision Conference*, pages 384–393, 2002.

[20] T. Mauthner, F. Fraundorfer, and H. Bischof. Region matching for omnidirectional images using virtual camera planes. In *Proc. of Computer Vision Winter Workshop, CVWW*, 2006.

[21] Rui Melo, João Barreto, and Gabriel Falcão. Camera calibration and image distortion correction for superior visualization in medical endoscopy. *Submitted to Transactions on Biomedical Engineering*, 2011.

[22] Krystian Mikolajczyk. *Detection of local features invariant to affine transformations.* PhD thesis, INPG, Grenoble, July 2002.

[23] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

[24] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2):43–72, 2005.

[25] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

[26] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42:145–175, May 2001.

[27] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[28] Davide Scaramuzza, Friedrich Fraundorfer, and Marc Pollefeys. Closing the loop in appearance-guided omnidirectional visual odometry by using vocabulary trees. *Robot. Auton. Syst.*, 58:820–827, June 2010.

[29] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 127–144. Springer, 2006.

[30] Richard Szeliski. Image alignment and stitching: a tutorial. *Found. Trends. Comput. Graph. Vis.*, 2:1–104, January 2006.

[31] Reg Willson and Steven Shafer. What is the center of the image? *Journal of the Optical Society of America A*, 11(1):2946 – 2955, November 1994.

[32] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, MIR '07, pages 197–206, New York, NY, USA, 2007. ACM.

[33] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.