FCTUC

University of Coimbra

Faculty of Sciences and Technology

Department of Physics

Techniques for keypoint detection and matching between endoscopic images

António Miguel Lourenço

Coimbra, 2009

Techniques for keypoint detection and matching between endoscopic images

Advisor: Prof. João Pedro Barreto

Committee:

Prof. Dr. Miguel Morgado

Prof. Dr. Hélder Araújo

Prof. Dr. Paulo de Carvalho

Prof. Dr. João P. Barreto

A Thesis submitted for obtaining the degree of Integrated Master
in Biomedical Engineering

Department of Physics

Faculty of Sciences and Technology,
University of Coimbra

July 2009

To my friends, my family and my girlfriend, Margarida,
for always believing in me...

Acknowledgements

Like any investigation project, my master thesis had experienced ups and downs during its course. So, it is essential to express my gratitude to some people because of their support and assistance provided.

I want to thank to my advisor Prof. João Pedro Barreto, for all the support that he gave me. I greatly acknowledge him for the excellent supervision and for providing an stimulating and cheerful research environment to work in, his encouraging support and his way to always find time for discussions.

In second place, I want to thank to all my lab friends: Michel, Queirós, Cristóvão, Melo, José Roquette and Abed. Thanks for all the laughs. Thanks Michel, Melo and Queróis for our lunch conversations that were always so insctructive. Thanks Abed for your support and our brainstormings that sometimes were vital to go forward with my work.

Obrigado aos meus pais. Sem eles, nunca teria tido a oportunidade de estudar aqui. Obrigado por todo o esforço, paciência e dedicação para me poderem dar uma boa educação.

Thanks to my sisters, brothers-in-law and nephews! For the motivation and love that you always gave me when I went home feeling lost. Thanks for daily calls pushing me foward. *Obrigado Maria Miguel, por perguntares, todos os dias, quando eu iria a casa.*

At last, the most important and grateful acknowledgement goes to my girlfriend Margarida. She was the most important person during this year and without her I had never achieve my work with success. Thanks for stand by my side when I needed the most and for always give me strength to carry on.

Thanks to all!

Contents

List of Figures

List of Tables

Abstract

The detection and description of local image features is fundamental for different computer vision applications, such as object recognition, image content retrieval, and structure from motion. In the last few years the topic deserved the attention of different authors, with several methods and techniques being currently available in the literature. The SIFT algorithm, proposed in [2], gained particular prominence because of its simplicity and invariance to common image transformations like scaling and rotation. Unfortunately the approach is not able to cope with non-linear image deformations caused by radial lens distortion. The invariance to radial distortion is highly relevant for applications that either require a wide field of view (e.g. panoramic vision), or employ cameras with specific optical arrangements enabling the visualization of small spaces and cavities (e.g. medical endoscopy).

One of the objectives of this thesis is to understand how radial distortion impacts the detection and description of keypoints using the SIFT algorithm. We perform a set of experiments that clearly show that distortion affects both the repeatability of detection and the invariance of the SIFT description. These results are analyzed in detail and explained from a theoretical viewpoint. In addition, we propose a novel approach for detection and description of stable local features in images with radial distortion. The detection is carried in a scale-space image representation built using an adaptive gaussian filter that takes into account distortion, and the feature description is performed after implicit gradient correction using the derivative chain rule. Our approach only requires a rough modeling of the radial distortion function and, for moderate levels of distortion, it outperforms the application of the SIFT algorithm after explicit image correction.

Chapter 1

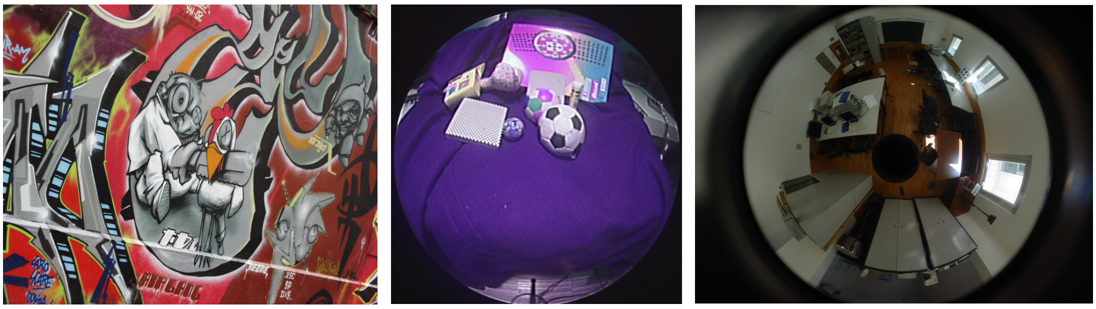
Introduction

The detection and extraction of local image features - henceforth called keypoints - is a low-level visual process. Such features must be stable in the sense that they can be detected and recognized across different views of the same scene. The usefulness of visual features is enormous and they are currently used in several computer vision tasks like motion tracking, visual recognition, 3-D reconstruction and camera calibration.

Different algorithms for extracting keypoints and visual features have been proposed in the last decades. The broadly used Canny edge detector [4] dates from the 80s and, since then, many other algorithms have appeared aiming to detect and match image features in a robust manner. Ideally these features must satisfy invariance properties, namely to changes in scale, rotation, illumination and viewpoint.

The Canny edge detector [4] and afterward the Harris corner detector [5], were invariant to image rotation and global illumination because they heavily relied on image gradients. In the 90s Lindeberg devoted a lot of attention to the problem of scale invariance [6]. He introduced in computer vision the concept of scale-space image representation and set the backgrounds for scale invariant feature detection and description.

With the Scale Invariant Features Transform, coined SIFT [7], David Lowe made a significant contribution for keypoint detection and description. His method, initially aiming object recognition, was carefully designed for being very stable and efficient. Lowe decided to describe the keypoints using histograms of image gradients computed in a neighborhood around the point. Before that point matching across views was usually performed by simply correlating local image patches around the extracted key-



(a) Viewpoint changes (b) Arthroscopic image (c) Catadioptric image

Figure 1.1: Images taken with different types of lens.

points. However, these methods were not robust because patches would change significantly under simple image rotations [8]. Several studies showed that SIFT descriptors are one of the most robust technique [8, 9] in terms of scale, rotation and minimal viewpoint invariance.

Most keypoint detectors and descriptors were designed for images captured by pin-hole cameras. However, the projection in many vision sensors that are broadly used in daily applications can not be described by the standard pin-hole model. Immersive environments, as well as surveillance systems, often require cameras with wide angle lenses, where the bending of the light rays when crossing the optics causes radial distortion. The distortion increases as we go far away from the center and is typically described by non-linear terms that depend on the image radius. The same problem arises when using cameras with mini-lenses or unconventional optical arrangements such as medical endoscopes.

Recently Hansen et al. [10] proposed an approach to extend SIFT for the case wide angle images. The method assumes that the calibration is known in advance, and that the images are back projected into a unit sphere centered in the effective camera viewpoint. Such transformation corrects the radial distortion and avoids that local regions surrounding a point undergo considerable changes under the action of pure rotation. A suitable scale space representation is obtained by solving the diffusion equation on the sphere. Unfortunately the approach is complex and computationally expensive because convolutions are performed in the frequency domain, which requires a spectral representation of the image using spherical harmonics [11].

1.1 Motivation

This thesis was developed in the context of the ArthroNav project, where the main goal is to perform computer aided surgery using as guidance the endoscopic video. Detection and matching of features in endoscopic video is problematic because the images present strong radial distortion (RD). Thus, our interest in detection and matching of features under RD was initially motivated by this problem. However, the usefulness of such research goes well beyond endoscopic imagery and medical applications. In fact, and as stated in the introduction, many vision sensors that are broadly used in daily applications can not be described by the standard pin-hole model. Examples are cameras with fish-eye lenses, often used in robotics and surveillance, and catadioptric sensors enabling panoramic imaging.

There is a broad variety of different methods and algorithms for detecting and matching keypoints across views [1, 2, 12]. The thesis focuses only on the SIFT approach originally proposed by David Lowe [2, 7]. This technique is one of the most robust in the literature, for which several improvements have been proposed [13, 14]. Nevertheless, and to the best for our knowledge, this is the first works proposing an extension of the SIFT framework to the case of images with RD. Closely related is the work of Hansen et al. [10]. They suggest back-projecting the image on an unitary sphere and building a scale-space representation that is the solution of the diffusion equation over the sphere. The keypoint detection is carried in the frequency domain using spherical harmonics. Such representation minors the problems inherent to planar perspective projection, enabling RD invariance and extra invariance to rotation. However, the approach requires perfect camera calibration and tends to be highly complex and computationally expensive. In contrast the method herein presented carries all the processing steps in the original image plane, only requires a rough modeling of the distortion function, and is a minor modification to the original SIFT algorithm that marginally increases the computation time.

1.2 Thesis Overview

This thesis is divided in four more chapters. In the next chapter the SIFT method is described in detail. In the third chapter, a performance evaluation of the SIFT under RD is presented. Section 3.1 introduces the so called division model for radial distortion [15, 16] that will be assumed through the remaining chapters. The data sets used for evaluating detection and matching, as well as the adopted metrics and criteria, are also explained. Chapter 4 derives the modifications to the SIFT algorithm that enable to improve detection and matching in images with RD, and discusses implementation issues with an impact on the computation time. The algorithm is compared against applying the original SIFT directly in images with distortion and in rectified images after RD correction. While the comparison in chapter 4 is carried over synthetically generated distorted images, chapter 5 provides an experimental evaluation on real imagery. In here, not only results with arthroscopic images are presented, but also results using conventional cameras equipped with lenses that introduce RD. Finally, conclusions about this work are reported.

The main contributions of the work can be summarized as follows:

- A thorough experimental evaluation of the performance of SIFT features in images with radial distortion. Although the invariance to scaling and rotation has been evaluated by previous authors [1, 8, 9, 13], to the best of our knowledge this is the first study that focuses into invariance of SIFT features to non-linear image distortion. It is provided experimental evidence that distortion strongly affects both detection and matching across different views.

- The experimental study is complemented by a theoretical interpretation of the observed results. This leads to a deep understanding about how RD impacts performance.

- A novel algorithm for feature detection under radial distortion. The key idea is to build the scale space representation using an adaptive gaussian filter. The time of computational time of the proposed method is close to the original SIFT over RD images, while the results of detection are improved.

- A novel algorithm for feature description is also explored. In here, a differential chain rule is used to invert the effect of the RD in the image gradients. This method allows to improve the feature matching performance for moderate amounts of non-linear deformation.

Instead of applying the standard SIFT method directly over the distorted images [17, 18], or perform image rectification to correct the distortion before running SIFT [19], the proposed approach adapts the SIFT to the distortion information. The final solution is a keypoint detector and descriptor that outperforms the standard SIFT algorithm directly applied over radial distorted images and that provides better results than image rectification for moderate levels of distortion.

Chapter 2

Scale Invariant Features Transform

2.1 The Multi-scale Approach: Scale-space Theory

Images of real world object and scenes when captured by a camera can vary under different conditions. The appearance of the same scene point in different images can substantially change with the viewpoint, the distance and the illumination conditions. The objective of a keypoint detector is to find points and features that can be easily recognized under varying image acquisition conditions. The scale and detail of a certain image feature across different views can undergo substantial variation. The scale-space theory allows to manipulate the image signal in a scale invariant manner by representing it at several levels of scale. The signal is successively low-pass filtered until it reaches an extrema (maximum or minimum) in scale which is the level where that particular signal structure should be handled.

The multi scale representation of an image is usually built by convolving the signal with a bi-dimensional Gaussian function (Eq.(2.1) and Eq.(2.2)). The filter standard deviation σ is the scale parameter/dimension. The key idea is to increasingly blur the image in order to obtain a measure of the signal variation in space. In fact, several results [20–22] prove that convolving the image with the Gaussian is the *canonical* way to obtain a scale-space representation of it.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.1}$$

$$L(x, y; \sigma) = I(x, y) * G(x, y; \sigma) \tag{2.2}$$

Given an image $I(x,y) : \mathbb{R}^2 \rightarrow \mathbb{R}$, its multi-scale representation $L : \mathbb{R}^2 \times \mathbb{R}_+ \rightarrow \mathbb{R}$, is defined as the solution of the heat diffusion equation (Eq.(2.3)):

$$\partial_\sigma L(x,y,\sigma) = \sigma^2 \nabla^2 L(x,y;\sigma). \tag{2.3}$$

with initial condition $L(x,y;0) = I(x,y)$ [23] and σ being the standard deviation of the filter.

As any formal theory, the scale-space is based in several axioms (see [23, 24]). Due to its properties, such as linearity, positivity and semigroup property, the Gaussian verify all the scale-space representation axioms. In addition the separability of the isotropic Gaussian filter is an attractive property that enables gains the computational time. Although this is not essential in terms of obtaining a suitable scale-space representation. A fundamental point when building the multi-scale representation is that the enhancement or creation of new structures is not allowed when smoothing the image from finer to coarse scale (increasing the scale parameter). The axiom of non-enhancement of new structures is meaningful in the sense that the structures with size smaller than σ must vanish away at level σ^2 of the scale-space representation, and can not lead to additional extrema at coarser levels of scale. This implies that at any level σ^2, a maximum could not increase and a minimum coud not decrease. Once more, the Gaussian verify this non-enhancement of local extremum property.

The scale-space framework in image processing relies on the *Principle of Scale Selection* stated by Lindeberg [24]: *In the absence of other evidence, assume that a scale level, at which some combinations of normalized derivatives assumes a local maximum over scales, can be treated as reflecting a characteristic length of a corresponding structure in the data.*

In other words, as we cannot predict the size (*how large*) of a certain feature, this framework provides a filtering of unknown structures just by convolving the image with a Gaussian at different scales and pick an extrema in Eq.(2.3). At each scale, the point that have a extrema at level σ^2 is extracted since it is illustrative of the correlation between the characteristic length of the signal feature and the standard deviation of the filter σ.

Considering the same image blobs represented at two different scales. The filtering process allows to have the same response in a scale invariant manner by convolving the images with the Gaussian and then solving the heat diffusion equation. Each image signal will be successively filtered until it reach a maximum in scale and starts to vanish in space and scale. This is the point where both structures should be handled since it is where they share more information with each other.

The scale-space framework is important in all steps of the most common and robust algorithms to extract keypoints. Since provides normalized responses (Eq.(2.3)) independent on the scale of the image, and allows for robust detection and description in the sense that all the computations required will be conducted in a scale invariant manner [6, 25].

2.2 Detection of Keypoints

The detection of points of interest, called keypoints in the SIFT framework, is the step where the most distinguishable points from an image are selected based on the local image properties. To be able to detect keypoints with high repeatability rates, a multi-scale approach, as introduced before, is necessary. Lowe [7] adopts a strategy that approximates the Laplacian of the scale-space representation of L by Difference-of-Gaussians (DoG). Let $I(x, y)$ be an image and $G(x, y, \sigma)$ a Gaussian function. The blurred version of $I(x, y)$ is obtained by its convolution with the Gaussian (Eq.(2.2)) and the DoG images are computed as:

$$DoG(x, y, k\sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{2.4}$$

In fact, the DoG is an approximation of the normalized Laplacian-of-Gaussians:

$$\frac{\partial L}{\partial \sigma} = \frac{L(x, y, k\sigma) - L(x, y, \sigma)}{k\sigma - \sigma} \tag{2.5}$$

$$\sigma \nabla^2 L = \frac{L(x, y, k\sigma) - L(x, y, \sigma)}{k\sigma - \sigma} \tag{2.6}$$

$$(k-1)\sigma^2 LoG(x, y, \sigma) = DoG(x, y, k\sigma) \tag{2.7}$$

The DoG approximates the LoG and provide the normalization required for scale invariant keypoint detection [24]. For instance, let consider a zoom change between the same scene. The normalized scale selection allows to absorb this scale difference

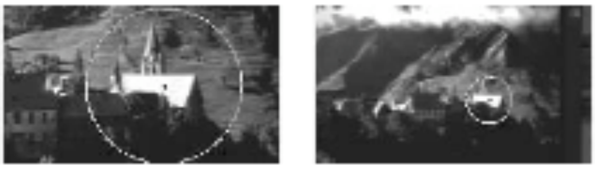


Figure 2.1: Support regions around the same keypoint at different scales [1]. The scale of selection is automatically scaled to cover the same region in the two images.

between them as it can be seen in Fig.(2.1).

The DoG pyramid strategy is intended to improve computational time and reduce the noise when computing the second order derivatives. In discrete data (as images), the derivatives are computed from pixel differences. If the first order derivatives are themselves susceptible to noise, the LoG is two times more sensitive even with the low-pass Gaussian blurring. Therefore the approximation DoG is attractive both in terms of noise and computational efficiency.

In Lowe's implementation, each group of 6 blurred images is called one octave (Fig.2.2). Its properties were studied in detailed to provide the most repeatable detection rate of keypoints. For each octave, 5 DoG images are produced to cover S=3 scales of keypoints selection but if more scales for keypoints selection are used the number of keypoints increases but the SIFT computation time also increases. The value of S=3 was experimentally obtained was the one that provides the better ratio between the number of keypoint detected and computational time.

The intervals of smoothness between images is defined by the number of scale to cover in each octave, and in Lowe implementation is of $2^{1/S}$. The keypoints are selected as being extrema of the DoG images in a 27 pixels neighborhood in the scale-space volume, as shown in the Fig.(2.3) [2].

After processing the first octave, the Gaussian image with standard deviation that doubles σ_0 (image blurring starts at $\sigma_0 = 1.6$) is down-sampled by a factor of two, taking every second pixel in each row and column. As the blurring of the image contributes to the creation of larger structures, the down-sampling will not contribute for aliasing. However, this assumption is only valid if is assumed that the image is enough blurred (one pixel in the original becomes two in the fourth smoothed version in each octave). At each octave, the first image starts to be sampled with $\sigma_0 = 1.6$. The down-sampling of the image by a factor of 2 is equivalent to increase the filter size being however a more computational efficient approach. The value of $\sigma_0 = 1.6$ was

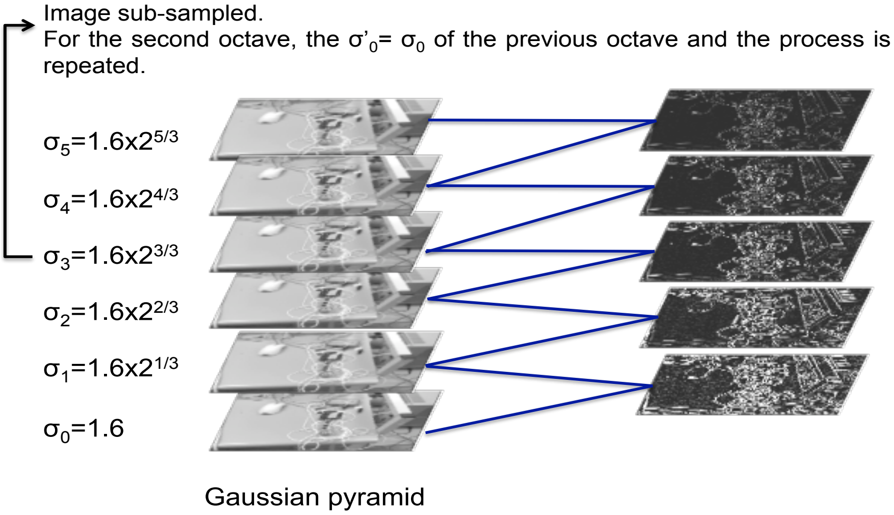


Figure 2.2: Scheme of the DoG's pyramid. 6 blurred versions of the original image and 5 DoG images are needed to filter 3 scales for each octave.

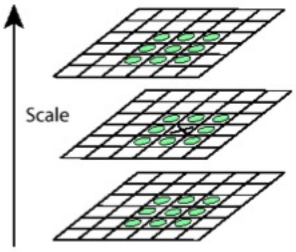


Figure 2.3: In DoG images each pixel is compared with its neighborhood of 3x3x3 pixels in the DoG images. The goal is detect points that are simultaneously extrema of $\frac{\partial L}{\partial \sigma}$ and where $\sigma^2 \nabla^2 L$ reaches a local maxima. A threshold is defined to select the most stable points and its value is equal to 0.04/S, if the pixel has lower intensity the point is discarded [2].

once more experimentally obtained as the one that shows more repeatability rates in detection [2].

To pick more extrema in the DoG images, the image size is doubled by bilinear interpolation. With the interpolation, the signal will be stretch which will create wider structures than in the original image. This effect will provide more keypoint candidates because small structures in the image will be bigger which allows to pick high frequencies of the image that the range of the DoG filtering does not consider in original size image.

When building the DoG pyramid, at a coarse scale the search for extrema would misestimate the spatial location of the feature. To avoid this issue, Lowe proposes a technique that will minimize the estimation errors of location [26].

2.2.1 Keypoints Location

In general, and due to bilinear interpolation, Difference-of-Gaussians pyramid produces too many keypoints and some of them belong to low contrast regions, or to non distinguisable edges, and therefore will not provide a reliable and consistent local descriptor for image retrieval applications. To avoid these issues, Lowe implemented some subpixel precision techniques that allow to correct errors of small shift on the location of the keypoints due to smoothing and eliminate non interesting keypoints.

Sub-pixel precision

For accurate keypoints location, Lowe considered a method proposed by [26] that allows a detailed interpolation of the points' scale-space coordinates (image coordinates and scale of detection) of the DoG. Using a Taylor expansion up to the quadratic term, it is given that

$$DoG(\mathbf{x} + \Delta\mathbf{x}) = DoG(\mathbf{x}) + \nabla DoG(\mathbf{x})\Delta\mathbf{x} + \frac{1}{2}\Delta^T\mathbf{x}\ H(DoG(\mathbf{x}))\Delta\mathbf{x} \tag{2.8}$$

where $\mathbf{x} = (x, y, \sigma)^T$. $H(DoG(\mathbf{x}))$ and $\nabla DoG(\mathbf{x})$ are, respectively, the Hessian matrix and the first derivative of the DoG image. The derivatives are computed using pixel differences of the smoothed neighborhood on a $3 \times 3 \times 3$ volume around the keypoints.

Differentiating to get the first order derivative it arises that:

$$H(DoG(\mathbf{x}))\Delta\mathbf{x} = -\nabla DoG(\mathbf{x}) \tag{2.9}$$

and the final offset, $\Delta\mathbf{x}$, is given by:

$$\Delta\mathbf{x} = -H(DoG(\mathbf{x}))^{-1}\nabla DoG(\mathbf{x}) \tag{2.10}$$

To the highest levels of the pyramid this interpolation becomes of vital importance due to the fact that the one pixel in those images correspond to a larger distance in the base image. If the offset $\Delta\mathbf{x}$ is higher than 0.5 in any dimension of the scale-space coordinates, the sampled point is moved to an adjacent position of the image and the accurate location is determined about this point.

Eliminating keypoints in low contrast regions

To eliminate keypoints in low contrast regions and avoid strong responses along edges, SIFT uses a Hessian matrix (Eq.(2.11)) that rejects unstable points based on the principal curvatures of DoG images. This Hessian is computed only in space and not in the scale-space volume like in the previous step.

$$H = \begin{pmatrix} DoG_{xx} & DoG_{xy} \\ DoG_{xy} & DoG_{yy} \end{pmatrix} \tag{2.11}$$

Since for SIFT proposes only an estimation of the ratio between the eigenvalues of Eq.(2.11) is needed, Lowe [7, 26] uses a metric based on the determinant and the trace of the Hessian and states that:

$$\frac{(tr(H))^2}{det(H)} = \frac{(DoG_{xx} + DoG_{yy})^2}{DoG_{xx}DoG_{yy} - DoG_{xy}^2} < \frac{(r+1)^2}{r}$$

being $r = \frac{\lambda_1}{\lambda_2}$, where λ_1 and λ_2 are, respectively, the largest and the smallest eigenvalue of Eq.(2.11). Since the Hessian is computed from pixel difference and r is a constant value, the verification if a point should be discarded or not is quite efficient. Once more, Lowe [2, 7] experimentally proves that $r = 10$ is the optimal value. If a lower value is selected more keypoints points are discarded along edges.

2.3 Description of Keypoints

After the detection of the keypoint and the refinement of their location, the next steps concern the computation of the final descriptor. In order to improve scale invariance, all the computations from now on are performed at the scale of selection of the keypoint in the Gaussian pyramid. This property is inherent to the use of the scale-space theory and it is one reason why the normalized responses across scales is so important.

The keypoints description is divided in two main steps. The first is to assign a characteristic orientation for each keypoints that will provide rotation invariance when computing teh keypoints descriptor. The descriptor step is where the local image information is encoded to be used for recognition and matching between scenes.

The the orientation and descriptor assignments to the keypoints are based on the magnitude and orientation of the image gradients at the scale of selection (Eq.(2.12) and Eq.(2.12)).

$$m(x,y) = \sqrt{(L(x+1,y;\sigma) - L(x-1,y;\sigma))^2 + (L(x,y+1;\sigma) - L(x,y-1;\sigma))^2} \tag{2.12}$$

$$\theta(x,y) = \tan\left(\frac{L(x,y+1;\sigma) - L(x,y-1;\sigma)}{L(x+1,y;\sigma) - L(x-1,y;\sigma)}\right) \tag{2.13}$$

2.3.1 Rotation Invariance : Orientation Assignment

The computation of the orientation is based on a histogram of the orientations built considering a neighborhood of 3σ, where is sigma the scale of selection of the keypoint. In practice, when computing a discrete approximation of the gaussian function, pixels at a distance of more than three times the standard deviation are small enough to be considered effectively zero. Thus contributions from pixels outside that range can be ignored. Each vote for account to the main orientation around the keypoint is computed using circular region to provide rotational invariance in the computation.

The algorithm to compute the orientation can be summarized as follows:

1. A histogram of 36 bins covering the 360 degrees around the keypoints is used.

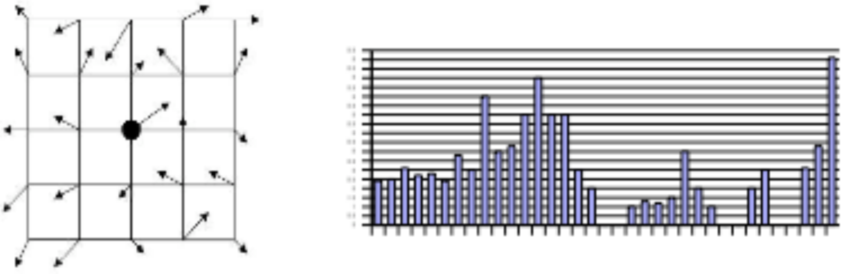


Figure 2.4: The black point represent the candidate point position after passing the elimination procedures. In the orientation histogram each bin represent 10 degrees covering the 360 degrees around the keypoint. Each bin holds the sum of all the magnitudes that count for it. Each sample is weighted by a Gaussian of 1.5σ to give less emphasis to contributions far from the keypoint.

The magnitude of each pixel contributes for the bin that covers it orientation.

2. A Gaussian weighting function with 1.5σ is used to give less emphasis to contributions far from the keypoint location.

3. The histogram of gradients could suffer the impact from the boundaries contribution between bins. This effect is reduced by convoluting the histogram with a Gaussian of 1.5σ. This minimizes the impact where a contribution could lye from one bin to another adjacent bin [9, 26].

Each keypoint could hold more than one orientation. In fact, any orientation within 80% of the main orientation is also assigned to the keypoint. In here, if this condition is verified, a new keypoint is created having the same scale-space coordinates but holding a different orientation that will contribute to have a different descriptor. This process greatly improves the robustness of the retrieval between images. Each orientation assigned contributes with different local information around the keypoint that could be successfully matched in another view [1, 2, 25].

2.3.2 Local Image Description: 128-dimensional Descriptor

This stage concerns building the descriptor that encodes local image information for posterior matching and recognition. SIFT descriptor is computed from a patch of 16×16 pixels around the keypoint after the rotation according to the main orientation assigned in the previous step. This provides the desired rotation invariance.

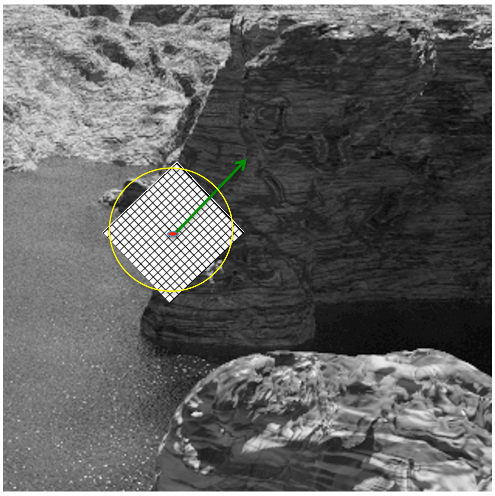


Figure 2.5: The patch used to compute the descriptor is rotated according to the orientation (green arrow assigned). The same procedure shown in Fig.(2.4) is used to account the values of the local magnitudes and orientations inside each 4 × 4 pixels subregions.

The final descriptor is built by dividing the 16×16 neighborhood in 16 subregions of 4×4 pixels. Each one providing 8 orientations which totalize the 128 components of the descriptor. This division in sub-regions allows to have pixels shifts until 4 positions in the image, while still contribute for the same sub-region [1, 2].

A Gaussian weighting function with standard deviation of one half the scale of the keypoint is applied to give less emphasis to the descriptor contribution far from the descriptor center. As, in general, SIFT produces too many keypoints and the descriptor window has a reasonable size, this technique is employed to give less emphasis to contributions that could lye very close to another keypoint descriptor. This technique increases the stability and the distinctiveness of 128-dimensional vector. To refine the elimination of boundary effects between each sub-region, interpolation is used to assign each gradients contribution to the correspondent subregion histogram [9].

The SIFT descriptor, besides the invariance to scale and rotation, also presents invariance to affine transformation and non-linear changes of illumination. By computing the descriptor from pixel differences, the changes in image brightness are discarded and, since it is normalized to the unit length, contrast changes are cancelled. To increase the robustness to non-linearities in illumination the descriptor is thresholded at 0.2 and then re-normalized [27]. This means that large gradients magnitudes are no longer very important. The value of 0.2 was experimentally determined by Lowe recurring to a large database of images [2].

2.3.3 Matching between Keypoint

Initially, Lowe [2,7] proposed to obtain matches by using the euclidean distance between a source and a target descriptor and considering a threshold of 0.8:

$$d_{euc} = \sqrt{\sum_{i=1}^{128}(d_1(i) - d_2(i))^2} < 0.8 \tag{2.14}$$

Nevertheless, in the last few years, the ambiguity distance gained more popularity because of its enhanced performance [10, 13]. The ambiguity distance instead of thresholding the euclidean distance between two descriptor, compare the distance between the closest and the second closest descriptor. So, considering d_2 the closest descriptor and d_3 the second closest descriptor , we can define the ambiguity distance as:

$$d_{amb} = \frac{\sqrt{\sum_{i=1}^{128}(d_1(i) - d_2(i))^2}}{\sqrt{\sum_{i=1}^{128}(d_1(i) - d_3(i))^2}} < 0.6 \tag{2.15}$$

This is the measure adopted in all the experiments in this thesis and has also been used in several studies in object recognition and image retrievel applications [1, 9, 13, 28, 29].

Chapter 3

SIFT evaluation under Radial Distorted Images

Several algorithms for invariant feature detection and matching have been proposed for images acquired by perspective cameras. However, the projection in many vision sensors that are broadly used in daily applications can not be described by the standard pin-hole model. Immersive environments, as well as surveillance systems, often require cameras with wide angle lenses, where the bending of the light rays when crossing the optics causes radial distortion. The distortion increases as we go far a way from the center and is typically described by non-linear terms that are function of the image radius. In feature detection techniques based on scale-space analysis, the image is represented at increasing scales by convolution with the Gaussian function. Given the set of scale-space images, scale-invariant features are found as local extrema of the first derivative in the scale dimension.

During the course of his PhD thesis, Mikolajczyk [1, 9, 25] published several results evaluating and comparing the most common used techniques for keypoint detection and matching under the more varied circumstances (scale and rotation, affine viewpoint changes, illumination changes and image compression). One of the contributions of this thesis is extending such evaluation for the case of images undergoing non-linear geometric deformations. We focus only in the SIFT algorithm, which proved to be one of most stable in the above mentioned comparative studies, [1, 9, 25], and we assume that the non-linear image deformation is radial distortion that can be fairly described using the first-order division model [15, 16].

3.1 The Division Model for Radial Distortion

The effect of lens distortion in image acquisition can be often described using the first order division model [15, 16]. Let $\mathbf{x} = (x, y)$ be a point in the distorted image I, and $\hat{\mathbf{x}} = (u, v)$ the corresponding point in the undistorted image $\hat{\mathsf{I}}$. The origin of coordinate system is assumed to be coincident with the distortion center, which is approximated by the image center [30]. The amount of distortion is quantified by parameter ξ (typically $\xi < 0$), and undistorted image points $\hat{\mathbf{x}}$ are mapped into distorted points $\mathbf{x}$ by function $\mathbf{f}$:

$$\mathbf{f}(\hat{\mathbf{x}}) = \begin{pmatrix} f_x(\hat{\mathbf{x}}) \\ f_y(\hat{\mathbf{x}}) \end{pmatrix} = \begin{pmatrix} \frac{2u}{1+\sqrt{1-4\xi(u^2+v^2)}} \\ \frac{2v}{1+\sqrt{1-4\xi(u^2+v^2)}} \end{pmatrix}, \tag{3.1}$$

The distorted image can be rectified using the inverse of distortion function :

$$\mathbf{f}^{-1}(\mathbf{x}) = \begin{pmatrix} f_u(\mathbf{x}) \\ f_v(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \frac{x}{1+\xi(x^2+y^2)} \\ \frac{y}{1+\xi(x^2+y)^2)} \end{pmatrix} \tag{3.2}$$

The function $\mathbf{f}$ is radially symmetric around the image center and its action can be understood as a shift of image points towards the center along the radial direction. The relationship between undistorted and distorted radius is given by :

$$\hat{r} = \frac{r}{1 + \xi r^2} \tag{3.3}$$

Radial distortion causes a space compression of the image information, which substantially changes the signal spectrum and introduces new high frequency components. To provide the notion of how much the image is compressed, we will often express the amount of distortion through the normalized decrease in the maximum image radius:

$$\%_{distortion} = \frac{\hat{r}_M - r_M}{\hat{r}_M} * 100 \tag{3.4}$$

with $\hat{r}_M$ and r_M denoting respectively the maximum values for the undistorted and distorted image radius.

Through this work we will always assume that image distortion follows the division model. However, most of the proposed methods and techniques can be easily extended to imagery affected by non-linear deformations that follow other types of parametric models (e.g. catadioptric distortion).

3.2 SIFT Detection and Matching in Images with Radial Distortion

The Difference of Gaussians (DoG) used during SIFT detection can be understood as a band-pass filtering action. The number of octaves of the Gaussian pyramid somehow define the range of image frequencies where the search for keypoints is performed. Since the radial distortion (RD) changes the image spectrum, it is natural that, when using a certain scale-space framework, the detection results are substantially different from the ones that would be obtained if the image had no distortion. This section tries to study both quantitatively and qualitatively the influence of RD during the detection stage.

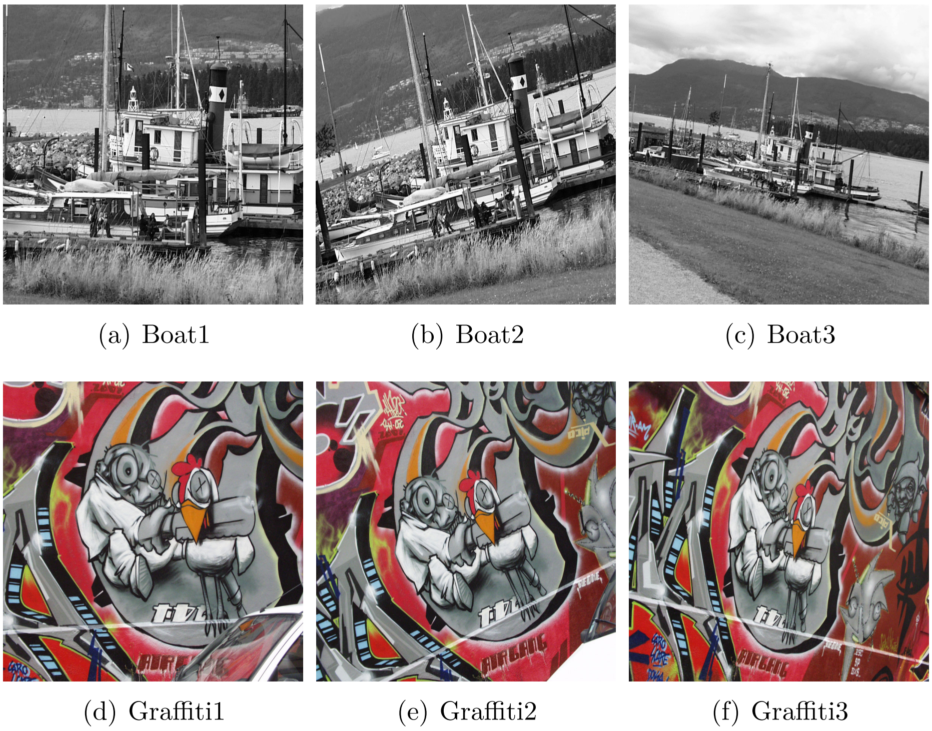
The evaluation of the SIFT detection under radial distortion is carried using two distinct data sets. The first data set comprises a set of images collected in the internet that covers different types of objects and scenes (Fig.(3.1)). The second data set is one of the benchmarks used by Mikolajczyk in his evaluation work [1,9,25] (Fig.(3.2)) [1]. These images are related by a known homography that enables to check the correction of SIFT detection and matching under rotation, scale and viewpoint changes.

The images of the two data sets are synthetically distorted assuming the division model presented above (Fig.(3.3)). It can be observed that, as the radial distortion increases, the image periphery is compressed. Since such compression tends to introduce high frequency components, it is to expect that detection will be strongly affected in the areas closer to image borders. It must also be taken into account that the total number of detected points is usually proportional to the image size. In general a big image will have more keypoints than a smaller one. Therefore, it is also to expect that the number of keypoints decays with the amount of added distortion because the size of the simulated image decreases.

In order to evaluate how distortion affects SIFT performance, we add radial distortion to the images of both data sets. The detection in each original image of data

[1] Available at http://www.robots.ox.ac.uk/ km

Figure 3.1: Data set 1 (DS1) : Images used for the synthetic experiments. The data set comprises a broad variety of scenes and visual contexts.



(a) Boat1

(b) Boat2

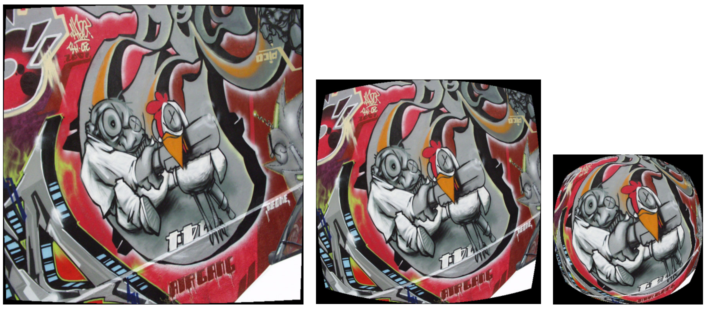
(c) Boat3

(d) Graffiti1

(e) Graffiti2

(f) Graffiti3

Figure 3.2: Data set 2 (DS2): The images in the top row are used to test scale and rotation. We will be referrer to them as *the boat sequence* (scales changes are of, approximately, 1.4 and 3.4 respectively). The images in the bottom row enable to test viewpoint changes (30% and 40%). This sequence will be named as *the graffiti sequence*. In each case, the normal scene (first image in both rows) is used to compare detection and matching against the other two.



(a) 5% (b) 25% (c) 55%

Figure 3.3: Syntheses of distorted images by artificially adding controlled amounts of radial distortion. As it can be seen the image compression in the periphery is much more noticeable than in the center (where it is assumed to be the distortion center).

set 1 is compared with the detection in its distorted versions. In addition the points are matched in order to evaluate the resilience of SIFT descriptors to radial distortion. A similar procedure was followed for the images of data set 2, but with detection and matching being performed between different views of the same scene. The experiments with data set 1 isolate the influence of radial distortion, enabling a thorough characterization of how it affects SIFT detection and matching. The experiments with data set 2 aim to measure the joint influence of radial distortion, rotation, scale and viewpoint change.

3.2.1 Ground Truth and Metrics for Evaluating Detection Performance

Consider an image of data set 1 and one of its distorted versions. Let S_0 and S denote the set of keypoints detected in the original and distorted images, respectively. The elements of S can either be points that have already been detected in the original image, or new keypoints that appear due to the high frequency components introduced by radial distortion. Henceforth, we will denote the former by S^d and the latter by S^{new} such that:

$$S = S^d \cup S^{new} \tag{3.5}$$

$$S^d = S_0 \cap S \tag{3.6}$$

The set S^d contains keypoints in the distorted image detected at a correct spatial location. However, and since SIFT detection produces normalized responses to scale, it is important to evaluate if these points are detected at the right scales. The correct assignment of scale is fundamental for achieving reliable matching across different views. Therefore set S^d is split in two subsets: S^c containing the points detected at correct scale and location, and S^{ws} being the set of points close in space but not in scale (detections at wrong scale).

$$S^c = S_0 \cap (S - (S^{new} \cup S^{ws})) \tag{3.7}$$

To decide if a point is detected at a wrong scale, first is checked if

$$\varepsilon_s = \left|1 - s^2 \frac{min(\sigma_{\mathbf{x}}, \sigma_{\hat{\mathbf{x}}})}{max(\sigma_{\mathbf{x}}, \sigma_{\hat{\mathbf{x}}})}\right| \tag{3.8}$$

where s represent the scale change between images (for the data set 1 its is assumed that $s = 1$) and σ denotes the scale of selection of the keypoints[2]. As in [1, 2, 13], we will allow a relative error in scale selection of $\sqrt{2}$. Since the scale is automatically normalized when the keypoint are detected, in practice the ε_s metric is defined by

$$\varepsilon_s = |1 - \frac{1}{\sqrt{2}}| \leq 0.3 \tag{3.9}$$

The repeatability in keypoint detection is evaluated using the following metric

$$\%_{\text{Repeatability}} = \frac{\#S^c}{\#S_0} * 100 \tag{3.10}$$

with # denoting the number of keypoints in each set.

The occurrence of new spurious detections due to radial distortion is quantified as follows:

$$\%_{\text{new detections}} = \frac{\#S^{new}}{\#S} * 100 \tag{3.11}$$

And finally the detection at wrong scale is characterized by the percentage of points detected at incorrect scale with respect to the points detected at a correct image location.

$$\%_{\text{Keypoints at Wrong Scale}} = \frac{\#S^{ws}}{\#S^d} * 100 \tag{3.12}$$

For the case of data set 2 the transformed image is distorted (zoom+rotation or

[2] $\hat{\mathbf{x}}$ and $\mathbf{x}$ denote the keypoints detected at the original and distorted images, respectively

viewpoint changes) and compared with the normal image (Fig.(3.2)). The ground truth S_0 comprises points that are simultaneously detected in a view pair, and satisfy the homographic relation between images up to a threshold value [2, 13]:

$$S_0 = (||\hat{\mathbf{x}} - \mathrm{H}\mathbf{x_t})|| < \sigma_{\mathbf{x}}) \tag{3.13}$$

where $\hat{\mathbf{x}}$ stands for the normal image keypoints, H for the homography between the normal image and transformed image (second and third image for both rows of Fig.(3.2)). The $\mathbf{x_t}$ represents the keypoints in the transformed image. The considerations and metrics discussed above for data set 1 are applied in the same manner.

3.2.2 Detection Results

This section presents the detection results in images with radial distortion. In order to build the repeatability curves, the RD is varied from 5% til 65%, and for each case the metrics introduced above are computed. From experimental results to evaluate detection, three main conclusions are drawn:

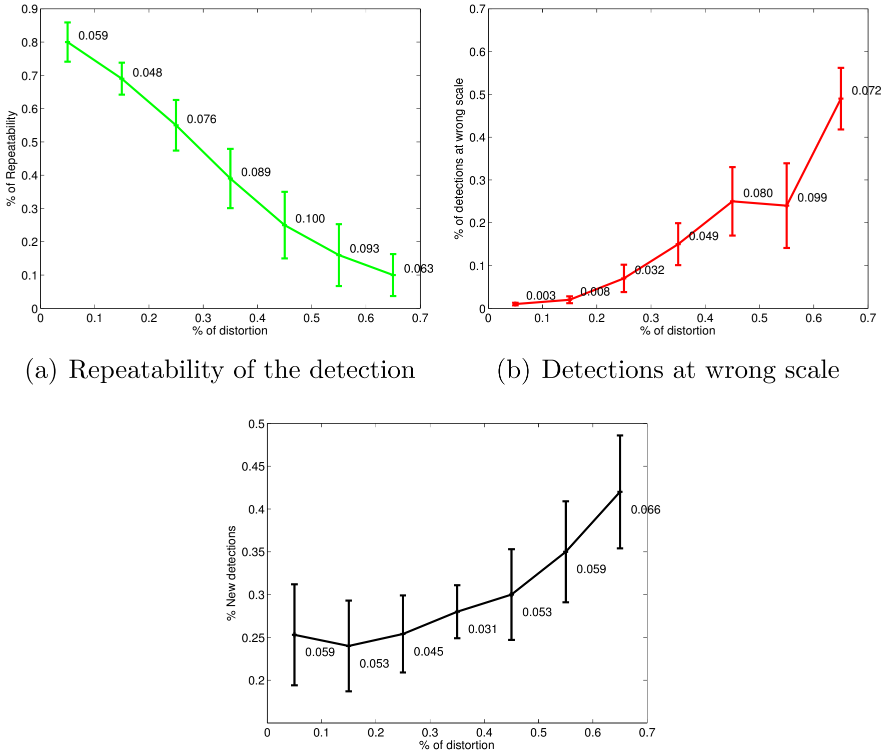
1. The repeatability of correct keypoint detection always decrease with increasing amounts of distortion.

2. New detections are more incident for higher values of distortion.

3. Changes in the scale selection of the keypoints are more incident for higher values of distortion.

In all the cases the repeatability presents the same behavior. For lower values of distortion ($\approx$ 5%), the repeatability does not decrease significantly but this effect becomes more pronounced as we increase the value of distortion. We observed that the keypoints detection in images with RD occurs at a finer scale than in the original image. This is a consequence of the compressive effect induced by the distortion. Since image structures are compressed they will provide a scale-space extremum at a lower level of the DoG pyramid.

The DoG range of filtering is constant all over the image, which leads to the lost of keypoints at the first levels of the pyramid in RD images. By the experimental results obtained and by the intuition of how the RD affects the image structures, we

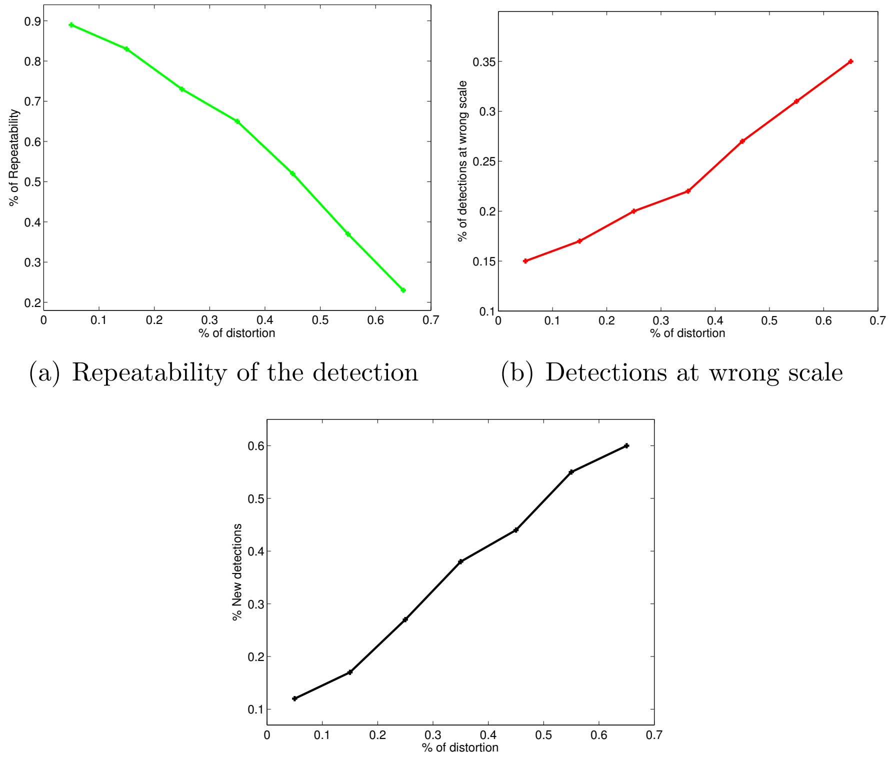


(a) Repeatability of the detection

(b) Detections at wrong scale

(c) New detections

Figure 3.4: Data set 1 detection results. In these cases, as several images are used we average the results and express the standard deviation in the data for each percentage of distortion considered. The green line stands for the correctly detected keypoints, the red for the points detected at wrong scales, and the black for the new detection, that were not present in the original image.



(a) Repeatability of the detection

(b) Detections at wrong scale

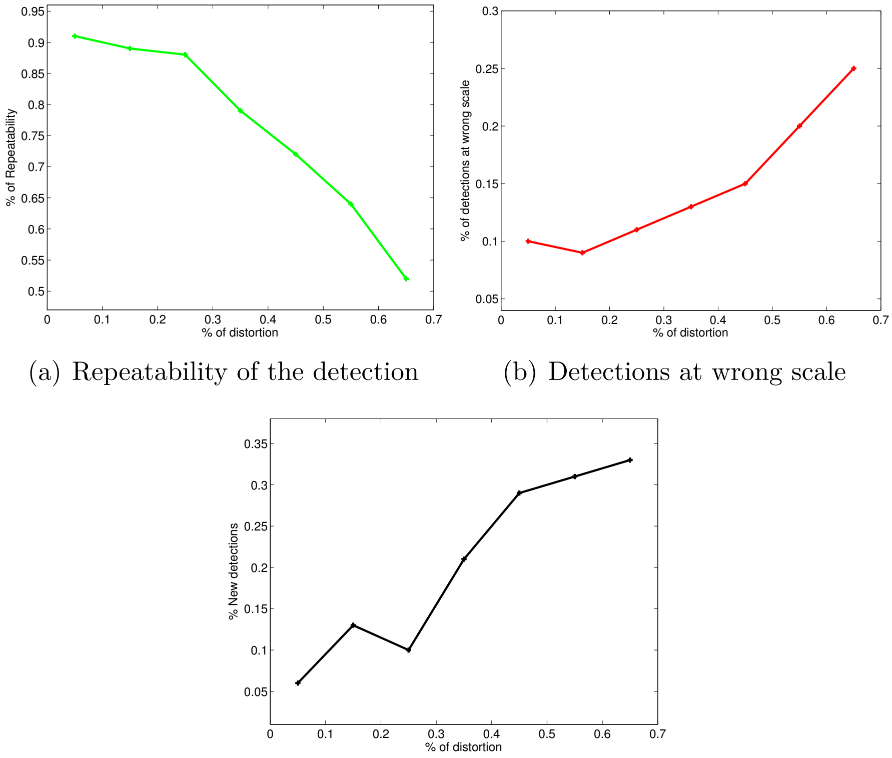
(c) New detections

Figure 3.5: Boat1-Boat2 detection results. This images shows a small zoom change and a considerable rotation. It can be seen that the new detection start to increase from a lower value than in data set 1. The small scale change between views allows that the image remains resilient to lower amounts of distortion, since the border of the image is not present in the original one.

can conclude that the lost of keypoints is more incident on those that appear at the lower scales than those detected at coarse levels of scale. In fact, keypoints detected at higher levels of the pyramid start to decrease their value of scale rather than disappear. They only vanish if their neighborhood is sufficiently compressed to not be considered as a keypoint by the SIFT detector.

The boat1-boat3 case confirms all the conclusions taken from the detection results analysis. This case presents a severe scale change, and the RD effect is only noticable when the common parts between boat 3 and boat1 are affected by RD. The detection at wrong scales and new detections rates start to be meaningful when the repeatability of the detection starts to be affected.

For the data set 2, mainly in the graffiti sequence (Fig.(3.7) and Fig.(3.8)), is observed that the detections at wrong scale is not as effective as in data set 1. In fact, when the image is submitted to transformations as viewpoint changes, the SIFT

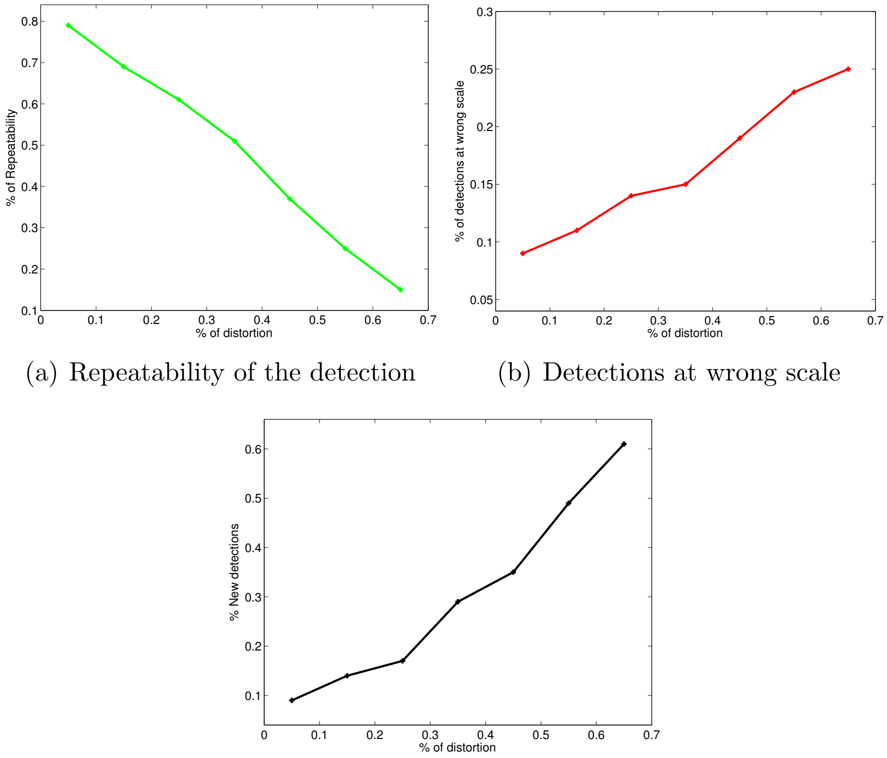


(a) Repeatability of the detection

(b) Detections at wrong scale

(c) New detections

Figure 3.6: Boat1-Boat3 detection results. The distortion only starts to affected the detection repeatability for moderate amount of distortion due to the severe scale change in this image.



(a) Repeatability of the detection

(b) Detections at wrong scale

(c) New detections

Figure 3.7: Graffiti1-Graffiti2 detection results. In here a scale change is also observed but the effect is minimum and is considered 1.4 [1].

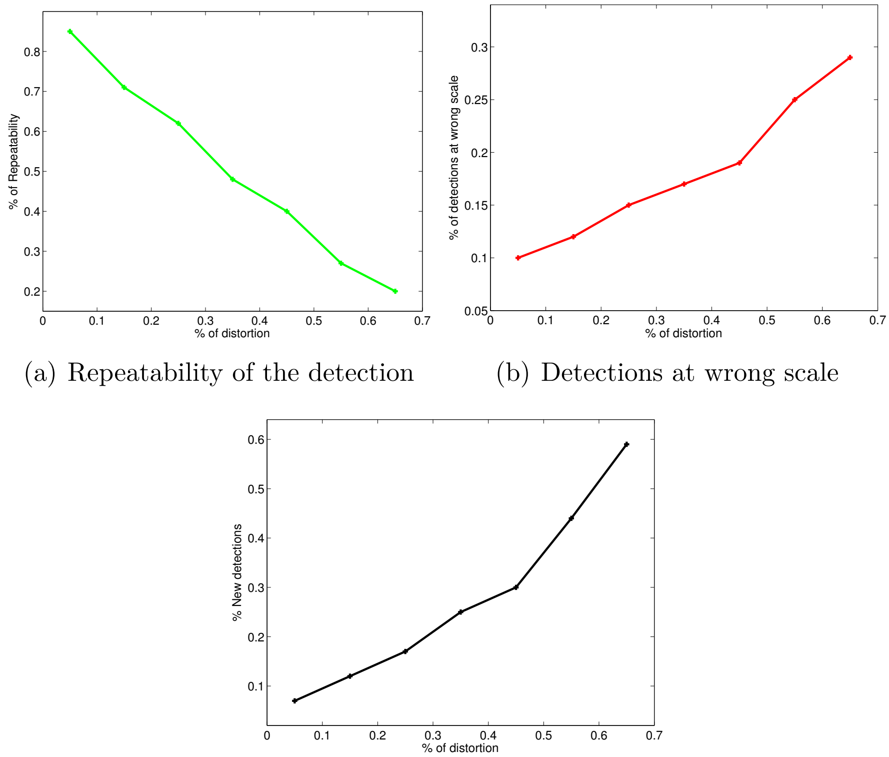
detector is invariant where the support regions around the keypoints does not change considerably. In these cases, the keypoints detected between views are detected at lower levels of smoothing, first and second octave of the DoG images. As mentioned before, in these levels of the pyramid is where the lost of keypoint correctly detected starts to occur due to the compressive effect of RD. As it can be seen in the experimental results, the detection at wrong scale does not reach the same values that in data set 1, where the goal is to test pure RD effect in the images.

The SIFT detector applies fixed size structures[3] to build the DoG pyramid for the detection Since the data in the periphery is more condensed, the Gaussian will pick contributions different from the undistorted image, that made the detection unstable (increasing number of newly keypoint detection).

We can conclude that the SIFT detector had two major problems that do not allow to have a same performance over RD images as in perspective images:

- At the first levels of the pyramid, since the Gaussian works with fixed size it does

[3] By fixed size structure it should be understood the size of the Gaussian defined to each scale of smoothness.



(a) Repeatability of the detection

(b) Detections at wrong scale

(c) New detections

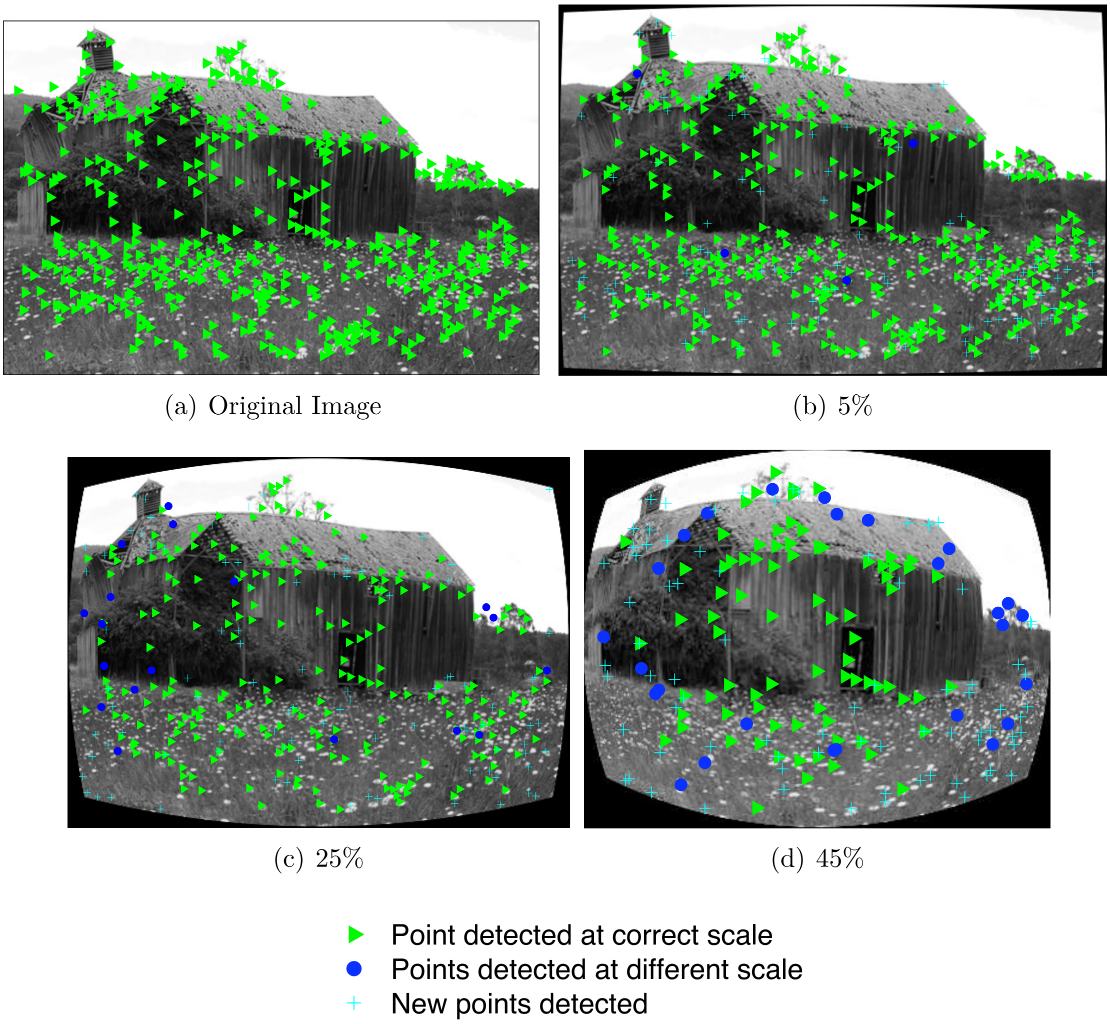
Figure 3.8: Graffiti1-Graffiti3. As in the other cases, the repeatability is highly affected by distortion. The scale change between images is of 2.4 [1].

not filter the compressed structures in the image periphery. The same reasoning is applied for the scale changes at coarse levels of the DOG pyramid.

- By doubling the size of the image prior to smoothing, the image structures are more wider so that DoG band-pass filtering can select them. Under radial distortion image, and due to the pixels shifting effect, nothing can ensures that the interpolation will be as stable as in the original image. This process can also introduce new detections that are not present in the original image.

3.3 Matching Evaluation under Radial Distortion

The SIFT computation of local image descriptors is based on image gradients computed at the scale of keypoint selection. The distortion shifts the image pixels towards the center along the radial direction. Such deformation impacts the image gradients and consequently the SIFT descriptors. This section aims to characterize this change in the descriptors that has direct consequences in the matching.



(a) Original Image

(b) 5%

(c) 25%

(d) 45%

▶ Point detected at correct scale
● Points detected at different scale
+ New points detected

Figure 3.9: Example of typical detection with distortion. The left side shows the original image to which was added increasing amounts of radial distortion. The marks represent the detected SIFT keypoints. The number of keypoints varies inversely with distortion mainly because of the decrease in image size/resolution. The green triangles correspond to points that are detected at the same scale both in original and distorted images. The detection repeatability is clearly affected by distortion with the effect being more pronounced in the image periphery. It is also interesting to observe that several keypoints in the original image are still detected in the distorted images at a lower scales.

The characterization is performed through two different experiments. The first experiment uses data set 1 and tries to isolate the influence of radial distortion in the descriptors. The detector is independently ran in the original undistorted image and in one of its distorted versions. The detected keypoints are matched and the obtained correspondences are split into correct and incorrect ones. A correspondence is considered to be correct if the image locations are consistent and the detection scales are the same. This means that the set of correct matches is always a subset of S^c defined in section 3.2.1. The second experiment uses data set 2 and studies the joint effect of distortion, rotation, scale and viewpoint change. In this case the base image is kept undistorted and it is matched against a distorted version of the second or third view.

3.3.1 Recall vs 1-precision Curves

In classification tasks is common the usage of ROC[4] curves for performance evaluation. This system of evaluation is based on sensitivity/recall against specificity. The recall provides the true positive ratio while the specificity gives the true negative rate. However, in this feature/image retrieval applications is hard to evaluate a true negative match between views. For this reason, instead of the usage of ROC curves, the recall *vs* 1-precision curves are used [1, 9, 10, 13]. The recall and 1-precision are given by :

$$recall = \frac{\#\text{correct matches}}{\#S^c} \tag{3.14}$$

$$1 - precision = \frac{\#\text{false matches}}{\#\text{false matches} + \#\text{correct matches}} \tag{3.15}$$

Finally this process is repeated varying the threshold of the Ambiguity distance from 0.4 to 1. As we increase the threshold, the number of keypoint descriptor that verify the ambiguity distance (Eq.(2.15)) also increases. To define the interval of thresholds considered, we decide to set the lower threshold as the one that observed at least one match (correct or wrong) in all the cases tested. As upper limit, we leave the threshold increase until all the keypoints in the original undistorted images had a match on the distorted image.

[4] Receiver Operating Characteristic

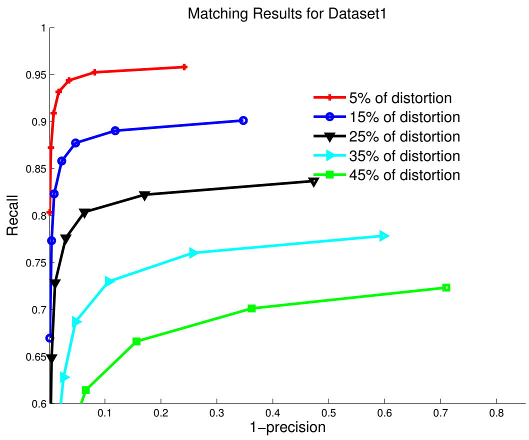


Figure 3.10: Matching performance evaluation for Data set 1. The effect of distortion deteriorates the effectiveness of the SIFT descriptor.

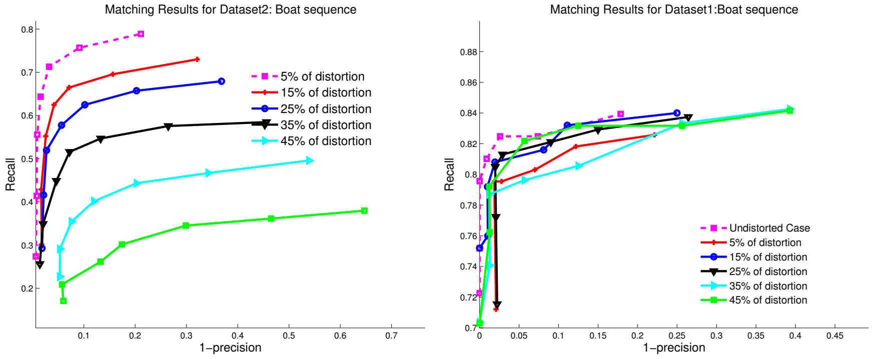


Figure 3.11: Matching performance evaluation for the boat sequence. Since the boat3 presents a severe scale change, the SIFT descriptor is not affected by RD.

3.3.2 Matching Performance Evaluation

To evaluate the SIFT descriptor performance independently of the detection under radial distortion effects, the distorted image is rectified and the SIFT detector is ran over this rectified image. The keypoints' descriptor are then calculated over the distorted image. This allows to see the effectiveness of the SIFT descriptor computed over the distorted image when the detection is more robust to distortion.

It can be observed from the performance evaluation curves that the SIFT descriptor is as much affected as the repeatability of the detection. The effect of the distortion over the descriptor is equally pronounced for rotation and viewpoint changes. When the image is rotated and then distorted, the pixel shifting according to the center of distortion lead to incorrect contributions for the descriptor that do not occur in the original or even in the undistorted rotated image. Generally, the SIFT behaves better for scale changes and small rotations, as in Boat3, than for viewpoint and severe

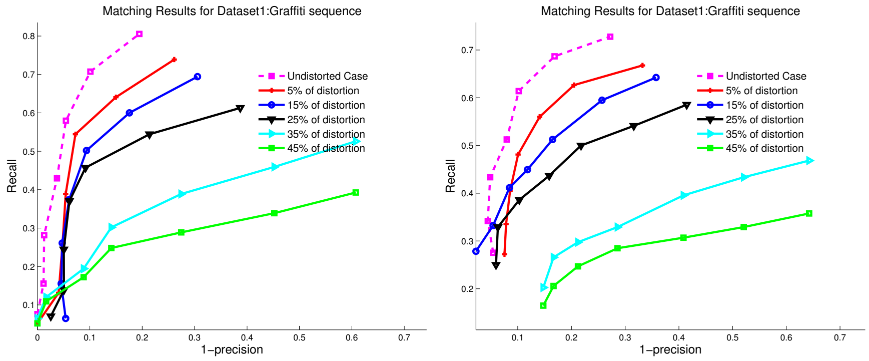


Figure 3.12: Matching performance evaluation for the graffiti sequence. In both cases, the behavior of the descriptor is highly affected. It must be regarded that SIFT only provide good results of minor affine changes in the viewpoint.

rotations changes [1, 2].

For the boat sequence it is observed that SIFT only starts affect the matching considerably when the common part between the views start to being affected. In the Boat1-Boat3 retrieval, the effect is partially imperceptible for low level of the distortion, the zoom effect is too severe and the descriptor is no really affected.

Since the descriptor is computed in a constant patch of 16×16 pixels, it is intuitively easy to understand that when the image is compressed the patch will have contributions that do not occur for the original image. Although the SIFT descriptor is robust to small shift in the histograms of each subregion of the descriptor, for higher amounts of distortion this influence starts to be noticed and the descriptor starts lose its performance. Another relevant constraint for the SIFT descriptor usage in RD images is that the effectiveness of the Gaussian weighting function starts to lose its effectiveness. As we increase distortion the contributions far from the keypoint tend to be more close to the keypoint. Since this effect increases with RD, the Gaussian function starts to give more emphasis to contributions that do not appear in the original image.

3.4 Conclusions and Perspectives

In this chapter was presented the SIFT performance over images with radial distortion and the following conclusions can be done:

- The repeatability of the detection and the detections at wrong scales are due to the compressive effect introduced by radial distortion. While the loss of repeatability is more pronounced at lower levels of the DoG pyramid, the wrong scale detections are more noticeable at coarse levels of scale.

- The compression induced by RD in the image spectrum leads to unstable detections that do not occur on the original image.
- The descriptor lose in performance also due to the shifting according to the center of distortion and because the window of the Gaussian weighting is maintained constant all over the image.

Since the DoG detection works with fixed size structures, the detection starts to fail where the features are more compress and are not considered in the range of the σ of the Gaussians. One fact that should be highlighted is that, under simulation, the size of the image decreases. Since the size of the image define the range of filtering of the SIFT, and hence the number of keypoint detected, SIFT is not the more appropriated approach in these cases.

One solution could be an adaptative Gaussians convolved with the image. In the next chapter will be derived a solution that allows to construct a more reliable scale-space representation of distorted images. Also the normalization of the scale space is not correct. Also the support regions around the keypoint for the orientation assignment should be automatically adapted to the size of the structure which will held to a more accurate orientation determination since the same similar contributions will be picked for the orientation in the original and in its distorted version.

It can be also observed that the descriptor is very affected by radial distortion. In fact, if keypoint are detected at correct scales, the deformation induces by distortion will not allow to match between view (data set 1 is the most clear example of this fact). Since the image deformation could be analyzed as an image warping, in the chapter 4 we will also explore a solution based on a differential chain rule, which models the distortion effect in the descriptor and that could be inverted.

Chapter 4

Detection and Matching Improvements

Up to now, we discussed the details of the SIFT algorithm and studied its performance in images with radial distortion. Such study allowed the identification of the main issues avoiding an efficient keypoint detection and matching. This chapter proposes several modifications to the original SIFT algorithm in an attempt to overcome the detected problems. We aim to come up with a method of minimal additional complexity that processes the images directly in the plane. We also take into account aspects of computational efficiency. The SIFT computation is already computationally expensive and the new approach must be carefully designed to be an efficient extension of SIFT for images with radial distortion.

4.1 Adaptative Gaussian filtering

A new approach for image adaptive blurring that generates a set of scale-space radial distorted images is introduced. An adaptive Gaussian filter that models the radial distortion effect in the image, is presented in this section. We assume that the division model provides a suitable description of the distortion and we base our derivations on it. However, the framework herein discussed can be easily adapted to any other parametric distortion model (e.g. catadioptrics).

4.1.1 Derivation of an adaptive filter

Let G_σ be a bi-dimensional gaussian function with standard deviation σ, $\hat{\mathsf{I}}$ denote the undistorted image, and I its corresponding distorted version. Like in section 3.1, the origin of coordinate system is assumed to be coincident with the distortion center, which is approximated by the image center [30]. The blurred image $\hat{\mathsf{L}}_\sigma$ is obtained by convolving the undistorted image $\hat{\mathsf{I}}$ with the Gaussian kernel (Eq.(2.2)). The brightness at pixel (s,t) is given by

$$\hat{\mathsf{L}}_\sigma(s,t) = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} \hat{\mathsf{I}}(u,v)\, \mathsf{G}_\sigma(s-u, t-v) \tag{4.1}$$

This is the standard convolution that SIFT performs for the case of the image being rectified to correct the distortion. In order to avoid interpolation and the computational burden of rectification, we aim to work directly with the distorted image I. Let $\mathbf{f}$ be the distortion function of Eq.(3.1) that maps the undistorted image coordinates (u,v) into the distorted image coordinates (x,y) and quantifies the amount of radial distortion by parameter ξ. It follows that

$$\hat{\mathsf{I}}(u,v) = \mathsf{I}(f_x(u,v), f_y(u,v)) \tag{4.2}$$

Replacing $\hat{\mathsf{I}}$ by I in Eq.(4.1) and performing a change of variables using the inverse distortion function (Eq.(3.2)), it arises that

$$\hat{\mathsf{L}}_\sigma(s,t) = \sum_{x=\alpha_-}^{\alpha_+} \sum_{y=\beta_-}^{\beta_+} \mathsf{I}(x,y)\, \mathsf{G}_\sigma(s - f_u^{-1}(x,y), t - f_v^{-1}(x,y)) \tag{4.3}$$

with

$$\begin{cases} \alpha_\pm & = \pm\frac{1}{\sqrt{-\xi}} \\ \beta_\pm & = \pm\frac{1}{\sqrt{-\xi}} \end{cases} \tag{4.4}$$

Let L_σ be the distorted version of the smoothed image $\hat{\mathsf{L}}_\sigma$. Changing the variables (s,t) by (h,k) using again the inverse distortion function, we can write

$$\mathsf{L}_\sigma(h,k) = \sum_{x=\alpha_-}^{\alpha_+} \sum_{y=\beta_-}^{\beta_+} \mathsf{I}(x,y)\, \mathsf{G}_\sigma(f_u^{-1}(h,k) - f_u^{-1}(x,y), f_v^{-1}(h,k) - f_v^{-1}(x,y)) \tag{4.5}$$

Replacing the $\mathbf{f}^{-1}$ by the expression of Eq.(3.2) and performing some algebraic manipulation, we finally obtain:

$$\begin{aligned}\mathsf{L}_\sigma(h,k) = \sum_{x=\alpha_-}^{\alpha_+} \sum_{y=\beta_-}^{\beta_+} \mathsf{I}(x,y)\, \mathsf{G}_\sigma \Big(& \frac{h - x + \xi r^2(h\delta^2 - x)}{1 + \xi r^2(1 + \delta^2 + \xi r^2\delta^2)}, \\ & \frac{k - y + \xi r^2(k\delta^2 - y)}{1 + \xi r^2(1 + \delta^2 + \xi r^2\delta^2)} \Big) .\end{aligned} \tag{4.6}$$

where

$$\begin{cases} \delta & = \sqrt{\frac{h^2+k^2}{x^2+y^2}} \\ r & = \sqrt{h^2 + k^2} \end{cases} \tag{4.7}$$

Note that now the smoothing kernel depends on (x, y) and (h, k) and Eq.(4.6) is no longer a straightforward convolution. However, if (h, k) is very close to the center, then $\xi r^2 \approx 0$ and the expression becomes a standard convolution. In this case there is no need to compensate the filtering for the distortion. If the pixels radius is far from the distortion center and the kernel is only significant in an area around the center of the convolution points (h, k), we have $\delta \approx 1$ and the Eq.(4.6) can be simplified in

$$\mathsf{L}_\sigma(h,k) \approx \sum_{x=\alpha_-}^{\alpha_+} \sum_{y=\beta_-}^{\beta_+} \mathsf{I}(x,y)\, \mathsf{G}_\sigma \Big(\frac{1}{1+\xi r^2}(h-x), \frac{1}{1+\xi r^2}(k-y) \Big) \tag{4.8}$$

This last equation is an approximation of the accurate adaptative filter (Eq.(4.6) (henceforth will be called the simplified adaptive filter). It represents a convolution by an isotropic gaussian filter whose standard variation decreases with the radius in order to adapt to the distortion. This isotropic approximation can be helpful in improving computational efficiency while building the scale space representation.

As mentioned earlier, with the derivation stated in Eq.(4.6) the response of the Gaussian filtering will be adapted to the distortion. For each convolution point is taken into account its distance to the center according to the distortion model adopt (see Fig.(4.1)). Using the simplification done in eq.(4.8), for a given scale σ_s, the filter compresses the convolution window from σ_s to $(1 + \xi r^2)\sigma_s$. The effect of this action increases while the filter moves far from the center of distortion. This new adaptive filter takes into account contribution of pixels closer to the convolution point than the naive filter when it is far from the distortion center. This is meaningful in terms of the intuition of how to efficiently blur a distorted image. Indeed, far from the center the

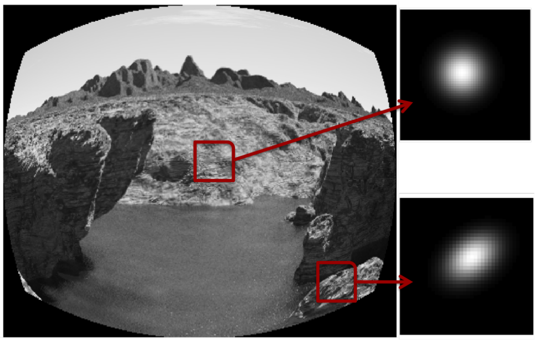


Figure 4.1: The Adaptative Gaussian Filter. It can be seen that as we increase the distance to the center of distortion the filter adjusts the radial distortion effect.

image contains more details. Then the convolution window needs to be compressed the to reduce the contribution of those new details introduced by the image structures compression.

The filter presented in Eq.(4.6) does not obey all the scale-space axiom because is missing the rotational invariance axiom. Nevertheless, in the literature some keypoint detector for affine transformation in the viewpoint relax this axiom in order to improve detection [8] but the scale of selection is normalized (iterative correction of the Gaussian to be isotropic) to compute the descriptor. However, the accurate filter is not rotational invariant, it allows to improve detection results.

For the simplified adaptive filter no precautions to build the multi-scale approach are needed since it is just a re-scaling of the σ of the naïve filtering who still obeys all the scale-space axioms.

4.1.2 Improving convolution time

In the initial stage of the implementation of the accurate adaptative filter, the gaussian blur as done recurring to a conjunct of pre-computed filters. This turns out the algorithm really slow, and the simplified adaptative filter becomes a more attractive approach.

$$\mathsf{L}_\sigma(h,k) \approx \sum_{x=\alpha_-}^{\alpha_+} \sum_{y=\beta_-}^{\beta_+} \mathsf{I}(x,y)\, \mathsf{e}^{\left(-\frac{(h-x)^2}{2\pi(\sigma(1+\xi r^2))^2} - \frac{(k-y)^2}{2\pi(\sigma(1+\xi r^2))^2}\right)} \tag{4.9}$$

$$\mathsf{L}_\sigma(h,k) \approx \sum_{x=\alpha_-}^{\alpha_+} \sum_{y=\beta_-}^{\beta_+} \mathsf{I}(x,y)\, \mathsf{e}^{-\left(\frac{(h-x)^2}{2\pi(\sigma(1+\xi r^2))^2}\right)}\, \mathsf{e}^{-\left(\frac{(k-y)^2}{2\pi(\sigma(1+\xi r^2))^2}\right)} \tag{4.10}$$

Since for each point of convolution the value of r is known, the kernel can be computed concerning the position of the pixel inside the image. In fact, ones can argue that since the kernel is computed in each position this does not speedup significantly the convolution. Although, it is possible to implement a speedup convolution that considered 5 multiplications at each time. This speedup convolution is also implemented in SIFT to speedup the blurring process.

4.1.3 Reducing the Number of New Detections

The evaluation of SIFT detection presented in section 3.2.2 shows that radial distortion introduces high frequency components in the image that lead to the detection of new keypoints. Such keypoints are spurious in the sense that they do not arise in the original undistorted images. Therefore, an ideal detector for images with radial distortion should be able to ignore such artifacts caused by the deformation. We discuss ways of achieving this objective.

Before blurring the image, the original SIFT starts by doubling the image size using bilinear interpolation, and then convolves the result with a gaussian kernel of standard deviation $\sigma_0 = 1.6$. This procedure tends to be unstable under radial distortion because the RD causes a shift in the pixel locations and there are no guarantees that interpolation will be carried with the same brightness values. In other words, the outcome of the interpolation step can have contributions that would not exist in the absence of radial distortion.

It must be regarded that the DoG detector compares the extremum in scale and space (Fig.(2.3)) and, when the image is distorted, pixels are highly compressed and sometimes vanish. One solution is to double the image after blurring (the same value of SIFT was chosen $\sigma_0 = 1.6$ to start filtering) and check the extremum in these double size version of the DoG images. The extrema are checked only in the pixels that are present in the original image rather than cover the all image pixels as in SIFT. This avoid the initial interpolation required by SIFT and the compression issues induced on the periphery of the image.

Since only the comparison for extremum selection is performed in the DoG, the blurred versions of the image that result from the convolution are kept at the same size. This is meaningful in the sense that all the orientation and descriptor computations are performed at the distorted image. The support region for the orientation computations is defined by the sub-scale of the keypoint : $(1 + \xi r^2)\sigma_s$. This also improves the descriptor rotation invariance computation under distortion.

4.2 Descriptor Improvement

Since the above solution for detection concerns a SIFT compensation that adapts to the distortion information and running the detection directly in the distorted space, a solution to provide a descriptor more resilient to radial distorted images is necessary.

4.2.1 Dimensionality Reduction of the Descriptor

Initially, a solution based on the representation of the descriptor in a sub-space of the 128 dimensions was tested. The descriptor vectors were analyzed to check if occurs a pattern on the number of dimensions affected to then discard them. For this, a PCA[1] analysis was done, recurring to the following procedure:

- The features were followed along 35% of distortion, incremented by 1% at each time. The features that were detected along all the distortion levels were stored to further analysis. 100 images were considered which totalize more than 3000 keypoints.
- Then PCA was applied in order to find a representation of the data covering 95% of the total variation of the descriptor.
- The number of meaningful dimensions according to the previous criteria were stored.

It was observed that any pattern occur in the data selected, e.g. the number of dimensions affected by distortion depends in the keypoints location inside the image and also from image to image. The principal reason for this fail is the non-linear effect introduced by the radial distortion in the image and by consequence in the image gradients.

[1] Principal Component Analysis is a standard technique for dimensionality reduction and has been used in computer vision tasks, namely in face recognition.

4.2.2 Implicit Gradient Correction in Radial Distorted Images

The SIFT descriptor is based on image first order derivatives at the scale of selection of the keypoints. Since the image mapping under radial distortion is modeled by Eq.(3.1), a solution based on a differential chain rule that models the gradients in the distorted image as being the same as in the original undistorted image was explored. The same reasoning applied in this section can also be made with any type of imaging systems, as long as they are described by a parametric models (e.g. catadioptric). The differential chain rule of the distorted model is derived considering the image first order derivatives. However, the implicit correction is performed at the scale of selection of the keypoint. This is done just for a sake of simplicity and due to the differential property of the convolution (Eq.(4.11)).

$$\partial_{x,y}(G * I) = G * \partial_{x,y} I \tag{4.11}$$

Applying the chain rule derivation on Eq.(4.2), we obtain

$$\begin{pmatrix} \frac{\partial \hat{\mathsf{I}}}{\partial u} \\ \\ \frac{\partial \hat{\mathsf{I}}}{\partial v} \end{pmatrix} = \begin{pmatrix} \frac{\partial \mathsf{I}}{\partial x}\frac{\partial f_x}{\partial u} + \frac{\partial \mathsf{I}}{\partial y}\frac{\partial f_y}{\partial u} \\ \\ \frac{\partial \mathsf{I}}{\partial x}\frac{\partial f_x}{\partial v} + \frac{\partial \mathsf{I}}{\partial y}\frac{\partial f_y}{\partial v} \end{pmatrix} = \mathsf{J} \begin{pmatrix} \frac{\partial \mathsf{I}}{\partial x} \\ \\ \frac{\partial \mathsf{I}}{\partial y} \end{pmatrix} \tag{4.12}$$

Using J as the jacobian of division model $\mathbf{f}$ (Eq.(3.1)), it follows that

$$\mathsf{J} = \frac{2}{k(1+k)^2} \begin{pmatrix} k(1+k)+4\xi u^2 & 4\xi uv \\ 4\xi uv & k(1+k)+4\xi v^2 \end{pmatrix} \tag{4.13}$$

with $k = \sqrt{1 - 4\xi(u^2 + v^2)}$.

Since $\hat{\mathbf{r}}$ is given by the Eq.(3.3), k could be express in terms of distorted radius by

$$k = \frac{1 - \xi r^2}{1 + \xi r^2}. \tag{4.14}$$

Replacing $\hat{\mathbf{x}} = (u, v)$ by $(x, y) = \mathbf{f}(\hat{\mathbf{x}})$ in the Jacobian expression and performing a suitable algebraic manipulation, it is possible to obtain J as a function of the distorted coordinates

$$\mathsf{J} = \frac{1 + \xi \mathbf{r}^2}{1 - \xi \mathbf{r}^2} \begin{pmatrix} 1 - \xi(\mathbf{r}^2 - 8x^2) & 8\xi xy \\ 8\xi xy & 1 - \xi(\mathbf{r}^2 - 8y^2) \end{pmatrix} \tag{4.15}$$

Instead of rectify the image of distortion and compute the gradients, this compensation is be applied over the distorted image. The SIFT descriptor computation will be modified to take into account the derivations above with a rough estimation of the distortion.

4.3 Results: Detection and Matching

Considering the derivations above, the detection can be performed using two different filters. For an evaluation of both the simplified and the accurate adaptative filters, we will consider the comparision with naïve SIFT directly over RD images and after rectify the image from distortion. The latter is most common approach for detection of keypoint under radial distorted images.

The evaluation of the descriptor for feature matching between images must be done using always the same detector. In order to fulfill this requisite, we choose to detect over the rectified image and then compute three types of descriptor for each feature:

- SIFT descriptor computation directly in the distorted image using original SIFT descriptor.
- Descriptor computation directly over the distorted image but with the proposed compensation for the gradients.
- SIFT descriptor computation on the rectified image.

This approach was only chosen because the rectification is the most stable case for detection being as well the common approach under RD situations. This condition is just to have the same number of positives (points detected at right scale and physical location) for the 3 test cases.

The same metrics and considerations of chapter 3.4 that were applied before are considered here. In this section, the result analysis will be divided by data set 1 and data set 2.

4.3.1 Data set 1

From the Fig.(4.2) it can be observed that, our filters overcome the problem of distortion for moderate amount of distortion. In all the cases, the accurate and the simplified adaptative filters show the highest scores for distortion until 25/35% of distortion. This is the boundary where our method start to lose in performance comparing with the rectification from distortion.

The derived filters for detection contributes with sub-scale approach inside each scale of filtering. Their properties allows to overcome the two main limitations of SIFT under RD. For the initial octaves of the scale pyramid, the Gaussian filter allows to detect points that SIFT does not consider anymore. The detection at wrong scales is also minimized since the adaptative Gaussians provides a more discretized search between adjacent scales at coarse levels of the pyramid.

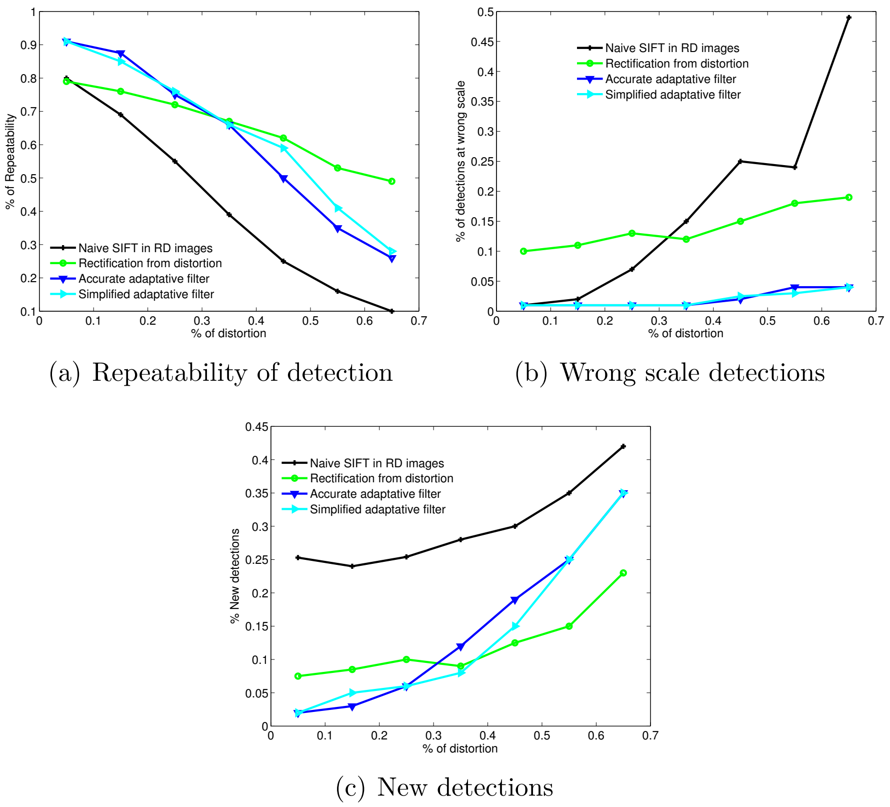
While we increase distortion, the image signal is tighted. Since a certain structure of the image is enough compressed to vanish, it is impossible to be selected anymore. So it is expected that for higher levels of distortion the rectification outperforms our adaptative filters. Nevertheless, our approach presents an improvement of detection until $\approx 35\%$ of distortion.

From the experimental evidence it is clear that for lower levels of distortion the method of implicit gradient correction outperforms the classic approaches, Fig.(4.2). It is observed that the image rectification is more valid approach for higher levels of distortion. Nevertheless, our method always provides better matching results when comparing with the use of SIFT directly in distorted images. The gradient correction technique allows to minimize the effect of the pixels shifting for moderate amounts of radial distortion. The region around the keypoints for compute the main orientation and the descriptor is also modified to perform the computations considering $(1+\xi r^2)\sigma$ instead of σ. This allows to have similar contributions in the distorted and in the original undistorted image, and then improve the descriptor resilience.

4.3.2 Data set 2

For the data set 2, it can be observed that the same improvements as before are achieved under common image transformations. In this section only the particular cases of boat1-boat2 and graffiti1-graffiti2 are presented. The two others can be seen in Appendix A.

The SIFT detector is only invariant to moderate amounts of viewpoint change when the region around the keypoint does not suffer significant changes. This occurs

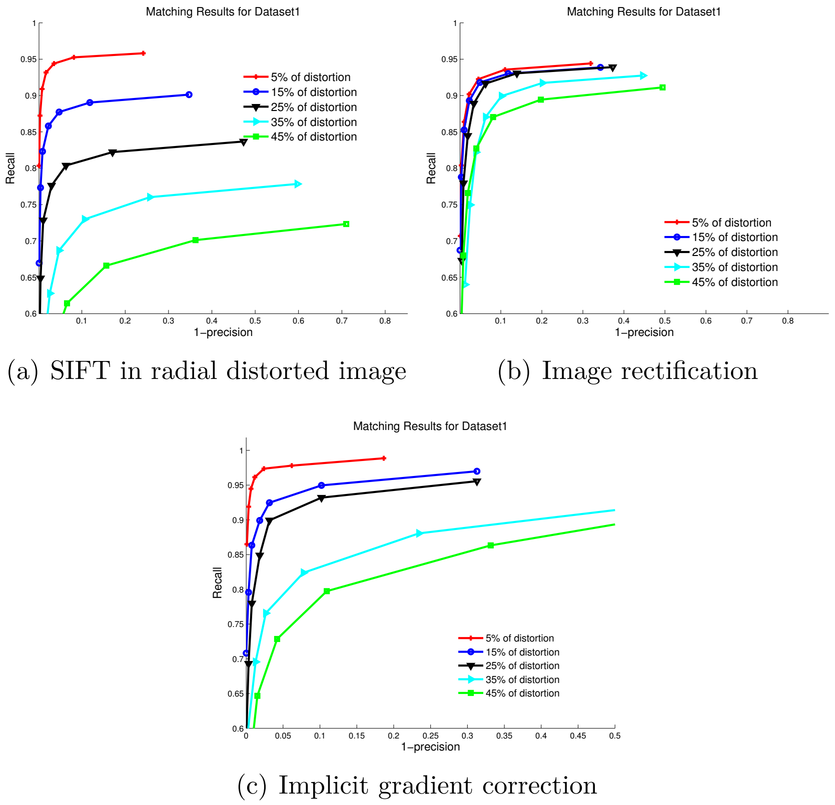


(a) Repeatability of detection

(b) Wrong scale detections

(c) New detections

Figure 4.2: Comparison of the 4 method for detection under radial distortion. Clearly the methods proposed outperforms the two standard approaches for moderate amounts of distortion.



(a) SIFT in radial distorted image

(b) Image rectification

(c) Implicit gradient correction

Figure 4.3: Recall curves. Our method is the most suitable approach at lower levels of distortion until (≈ 25% of distortion).

at the first levels of the pyramid (first and second octave of the DoG). In these cases, our method also outperforms the most common methods for moderate amounts of distortion due to the sub-scale property our filters provide (Fig.(4.6)).

In this data set, the wrong scale selection is not as evident as in data set 1. Since our method overcomes the detection at wrong scale, we can conclude that the small break in peformance of repeatability when comparing to data set 1 comes from the keypoints vanishing under radial distortion effects. That is also why the rectification allows better scores for higher amounts of distortion in this data set. The rectification is not so sensible to the keypoints vanishing effect. By interpolating the image to correct from distortion, the structures will be approximated to those present in the original image.

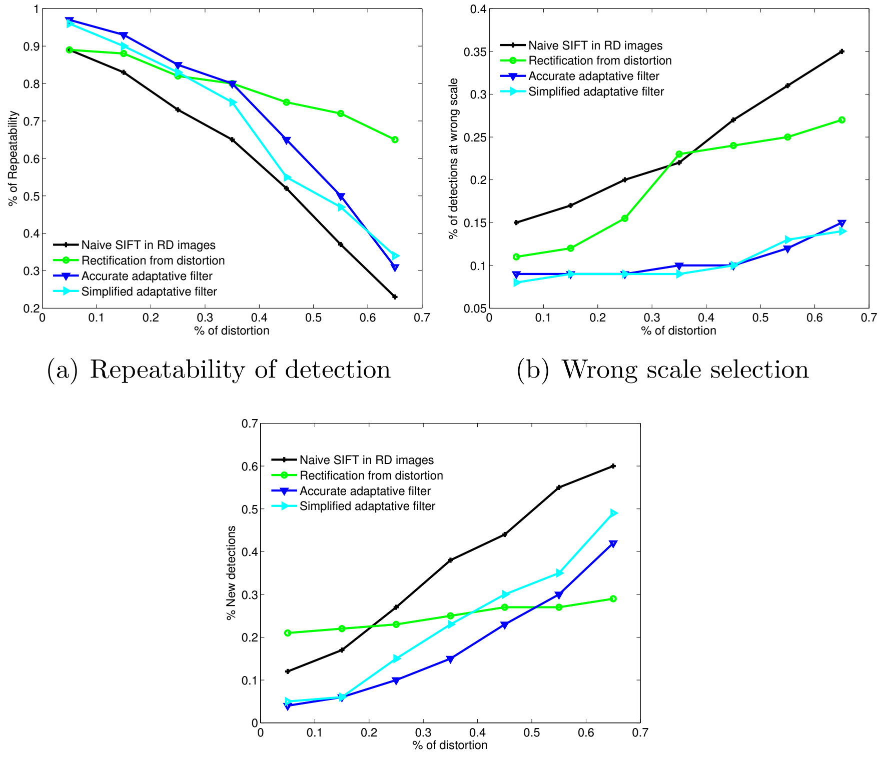
For the boat and graffiti sequences (Fig.(4.5) and Fig.(4.7)) it can be seen that the implicit gradient correction outperforms the other approaches until $\approx 25\%/30\%$. This is a very important fact that highlight the effectiveness of the implicit gradient correction while preserve the role of capabilities of the original SIFT descriptor.

Concerning the boat sequence, namely the boat1-boat3 case, it can be seen that the proposed method is capable of performing equivalently to SIFT even when the distortion is not that pronounced (the zoom is very severe and the common parts between views are not very affected by distortion). This confirms what was said in the last paragraph and also proves that since the proposed method do not require any interpolation of the image it does not deteriorate the descriptor performance.

We conclude that implicit gradients compensation over the distorted image is an effective way to compute the descriptor. The integration of the detector proposed with this new compensation for description will lead to an effective keypoint detector and descriptor under radial distortion.

4.4 Conclusions

In this chapter, the main contribution of this thesis was formally presented as well as experimental evidence of its performance. It can be observed that complex and simplified adaptative filters are the most suitable approaches for keypoint detection under moderate amounts of radial distortion. In all the cases, these filters always

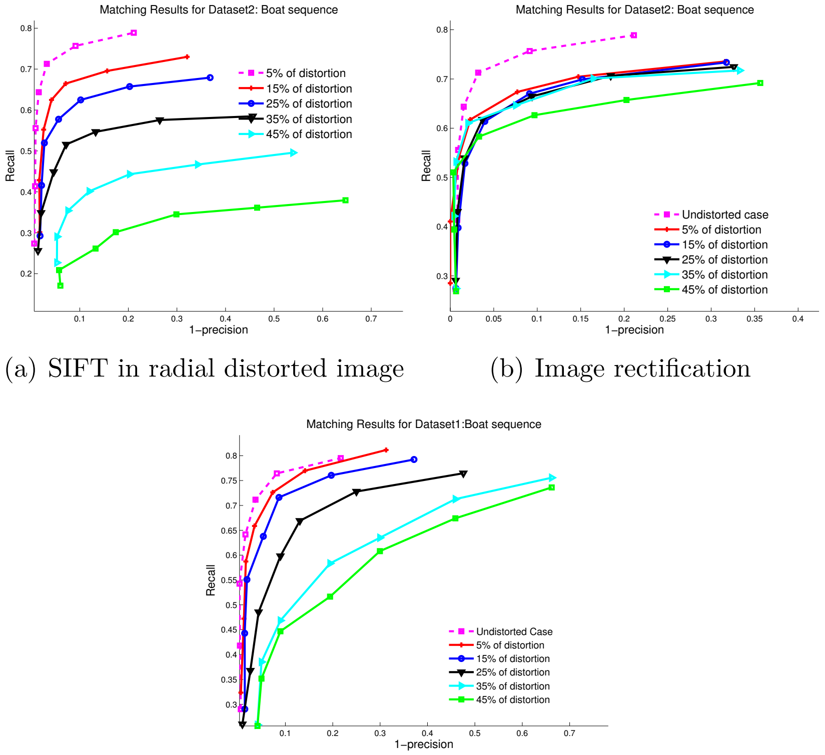


(a) Repeatability of detection

(b) Wrong scale selection

(c) New detections

Figure 4.4: Detection evaluation for boat1-boat2 case. For lower of distortion our method performs better than the two classic approaches.

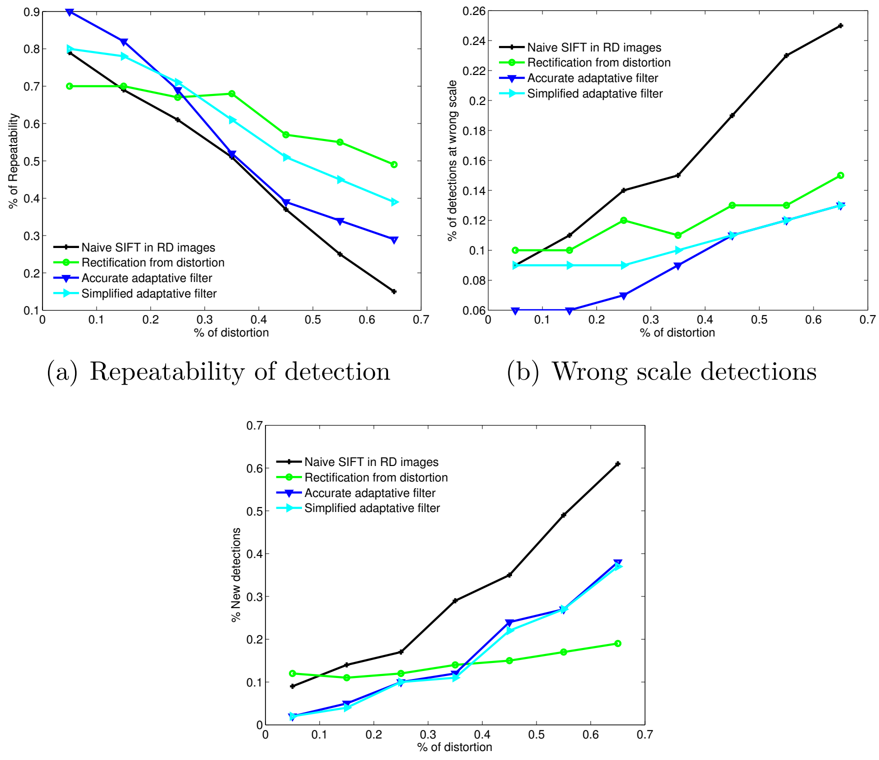


(a) SIFT in radial distorted image

(b) Image rectification

(c) Implicit gradient correction

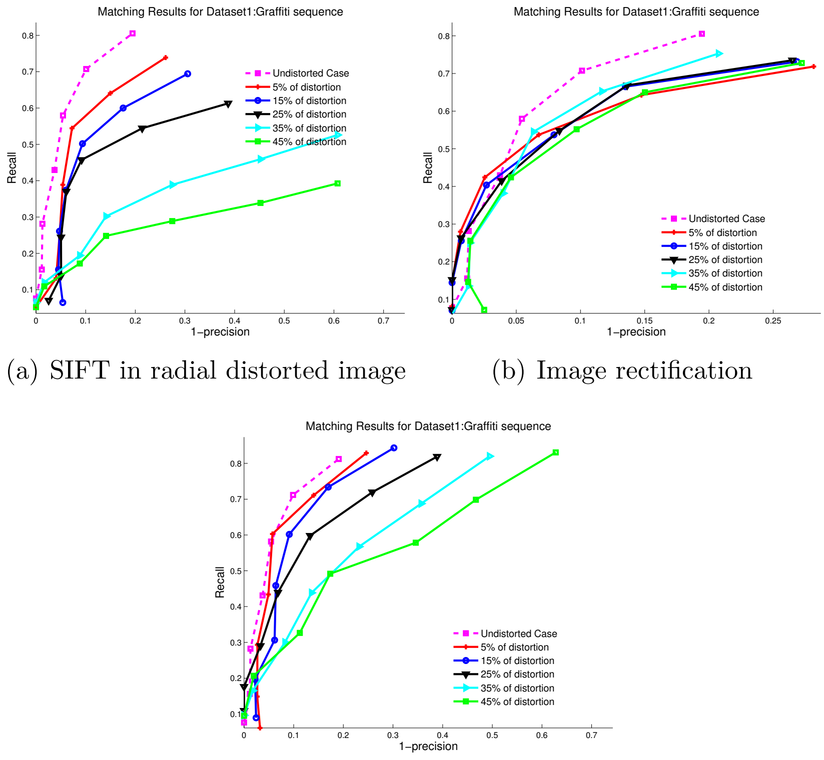
Figure 4.5: Recall curves for boat1-boat2 case. The method in (c) is the most suitable approach at lower levels of distortion (≈ 30% of distortion).



(a) Repeatability of detection (b) Wrong scale detections

(c) New detections

Figure 4.6: Detection evaluation for graffiti1-graffiti2 case. For lower percentage of distortion our method performs better than the two classic approaches.



(a) SIFT in radial distorted image (b) Image rectification

(c) Implicit gradient correction

Figure 4.7: Recall curves for graffiti1-graffiti2 case. The method in (c) is the most suitable approach at lower levels of distortion (until ≈ 30% of distortion.

outperform the naïve SIFT applied over radial distorted images and for moderate levels of distortion the image rectification.

The accurate filter shows the best results in some cases, but, at this point, the obvious solution is the application of the simplified filter. This filter obeys all scale-space axioms and no constraints are needed when using it. It also allows to speedup the processing of filter showing similar results to the accurate filter and, surprisingly, overcoming it for some cases. Nevertheless, the accurate filter could be optimized in order to fill the scale space theory by the projection of the filter to be isotropic. To the best of our knowledge, this is the first work that proposes to extend SIFT to RD images with all the processing being done over the image plane.

The gradient correction for the SIFT descriptor computation presents also a suitable approach. It performs better than image rectification and it is a suitable approach to integrate with the adaptative filters derived. However, for higher amount of distortion, the implicit correction starts to have the same effect has the original SIFT descriptor. One possible solution for higher values of distortion is the projection of the corrected gradients to the rectified image coordinates. This step will not require gradient interpolation and could be performed by the inverse distortion mapping (Eq.(3.2)) just to avoid the compressive effect over the gradients of the image.

Chapter 5

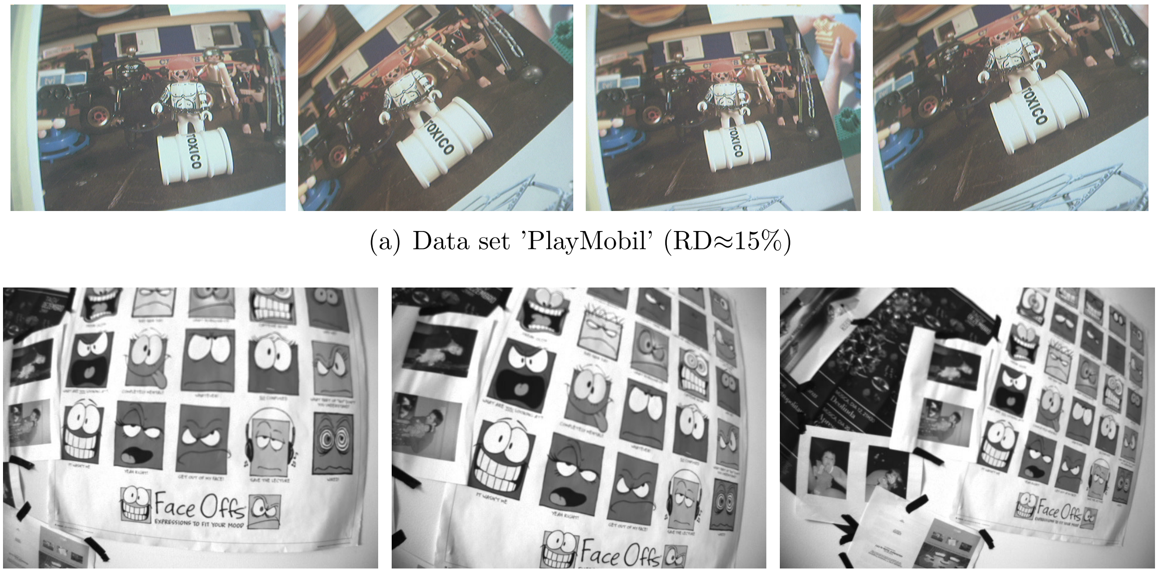
Experimental Real Cases

In this chapter it will be considered real case with images taken by cameras with lenses that cause radial distortion. First, we show the improvement in real case images taken from PointGrey cameras. In each case a comparison between the four methods for detection in study will be performed. Since in this case it is not possible to build repeatability curves, the results is expressed in terms of absolute values of keypoint detected between scenes. For the descriptor, the simplified adaptative filter will be used to extract the keypoint prior to the descriptor computation for the three types of descriptor. In here, the used lens present distortion of $\approx 15\%$ and $\approx 25\%$.

5.1 Real Images with Radial Distortion

Since the presented method shows under simulation better performance for lower levels of distortion, in here is considered two different datasets (Fig(.5.1)). One with less then 25% of distortion and other with more than this threshold where the method proposed performs worse than the rectification. For this, two data sets that consist of images of a textured planar surface were acquired. This means that every two images are related by an homography that enables verifying if the matches are correct and also to generate the ground truth between images. In order to do this, a rough estimation of the distortion parameter [3] to undistort the images is made. Then, the homography between the different image views was generated by hand using ten correspondences. The ground truth was generated by considering the criteria introduced in chapter 3.

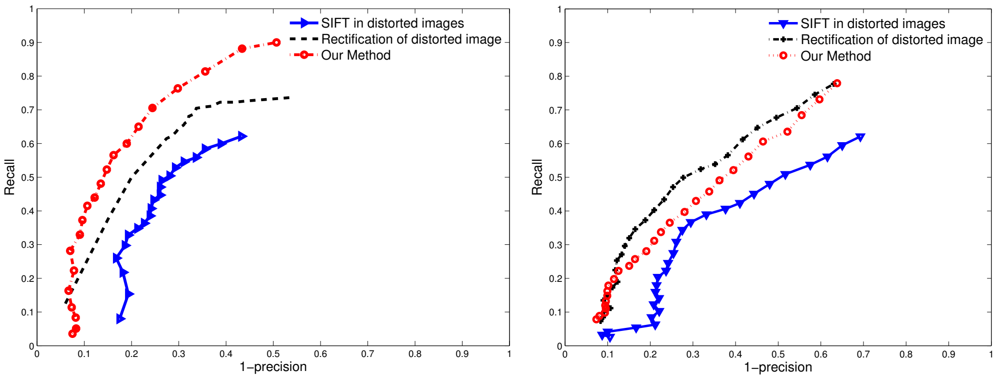
For the Playmobil data set, it can be confirmed that the method proposed perform



(a) Data set 'PlayMobil' (RD≈15%)

(b) Data set 'Smile' (RD≈25%)

Figure 5.1: The figure shows the two data sets employed in the real experiments. For each case the radial distortion was roughly estimated using the projection of lines [3].



(a) PlayMobil

(b) Smile

Figure 5.2: Curves of precision-recall for keypoint matching. Our method of implicit gradient correction shows the best performance for the PlayMobil data set. The images are taken from different viewpoints (Fig. 5.1(a)), which proves that our method overcomes the problem of radial distortion while preserving the invariance properties of the SIFT descriptors. Image rectification performs better for the Smile data set which was acquired by a camera with ≈ 25% of distortion. These results are in perfect accordance with those obtained under simulation.

Table 5.1: The Playmobil1 is taken as the ground truth. It can be seen that our simplified/accurate adaptative filters allows to have more points between scenes than the standard approaches.

Methods	PlayMobil1	PlayMobil2	PlayMobil3	PlayMobil4
Accurate \ Simplified adaptative filter	1234\ 1212	915\ 901	1195\1045	746\712
SIFT in RD images	1045	795	976	565
Image rectification	995	764	895	669

Table 5.2: The Smile 1 image is taken as the ground truth to compare with the other two cases. It can be seen that our simplified/complex filters allows to have more points between scenes than the standard approaches. It must be regard that the appearance between scenes changes severely and which is a challenge for all the approaches tested.

Methods \ Images	Smile1	Smile2	Smile3
Accurate \ Simplified adaptative filter	2394 \ 2285	2101 \ 1995	1205\ 1175
SIFT in RD images	2254	1598	875
Image rectification	2176	1675	975

better than the two classic approaches for small amount of distortion. The implicit gradient correction method beats the two other approaches both in recall and precision. From the smiles dataset, it can be observed that the image rectification performs better than our method. This is the limit where the implicit method starts losing in matching performance against rectifying the image from distortion and apply SIFT. Nevertheless, our method performs better than applying SIFT directly in the distorted image.

5.2 Applications to Medical Imaging

The images were acquired by a Smith&Nephew Camera 460H with lens used in medical interventions. Although the arthroscopic camera has a 576×720 pixels, the image is characterized to have a smaller useful area. To process the images only the useful area ($\approx 550 \times 520$ pixels) is taken into account. Our simplified adaptative filter with the implicit gradient correction is compared against SIFT over the arthroscopic images and image rectification from distortion. In images of this size, SIFT detects in average 800 to 1000 keypoints.

Up to here, all the results obtained with the PointGrey Cameras can be extended to the arthroscopic images assuming the first order division model for radial distortion.

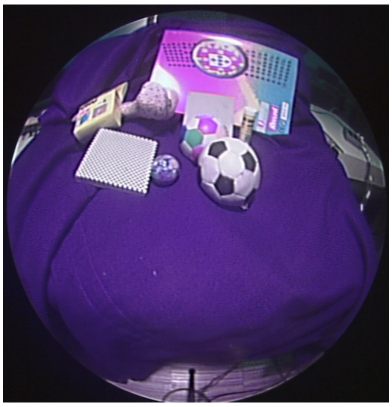


Figure 5.3: Scene acquired with an Arthroscopic Camera. The lens used induce strong radial distortion in the image.

In here, just some vizualization work will be explored.

We consider 10 arthroscopic images that could be consulted in Appendix B. For each image, three types of methods are used. So, the naïve SIFT will be ran in distorted images and rectified images, and the proposed simplified adaptative filter will be ran over distorted images where the descriptor incorporate the implicit gradient correction proposed. Since any pattern will be followed in here, and just the intuition of the performance under arthroscopic images should be perceived, the image Fig.(5.4) will be taken as basis. No selection of the positives detection between views is done and the maximum number of corrected matches is selected by RANSAC[1]. To estimate the maximum number of inliers (correct matches) a online function provided by Peter Kovesi was used [31]. This function allows a robust fit of the fundamental matrix from a initial set of correspondences between images, which allows to check the total number of matches between views. To evaluate that the proposed approach is viable, we show a comparative graphic of the computational time of each approach. For the rectification of the images, also the step of distortion correction is incorporated in the time evaluation (Fig.(5.5)).

Surprisingly, the rectification does not provide a substantial improvement in terms of detection. Nevertheless, it supplies more correct matches between the different views than the SIFT over RD images.

The proposed method allows to have more detected points in all the views. Also the average number of corrected matches is higher which shows the effectiveness of

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

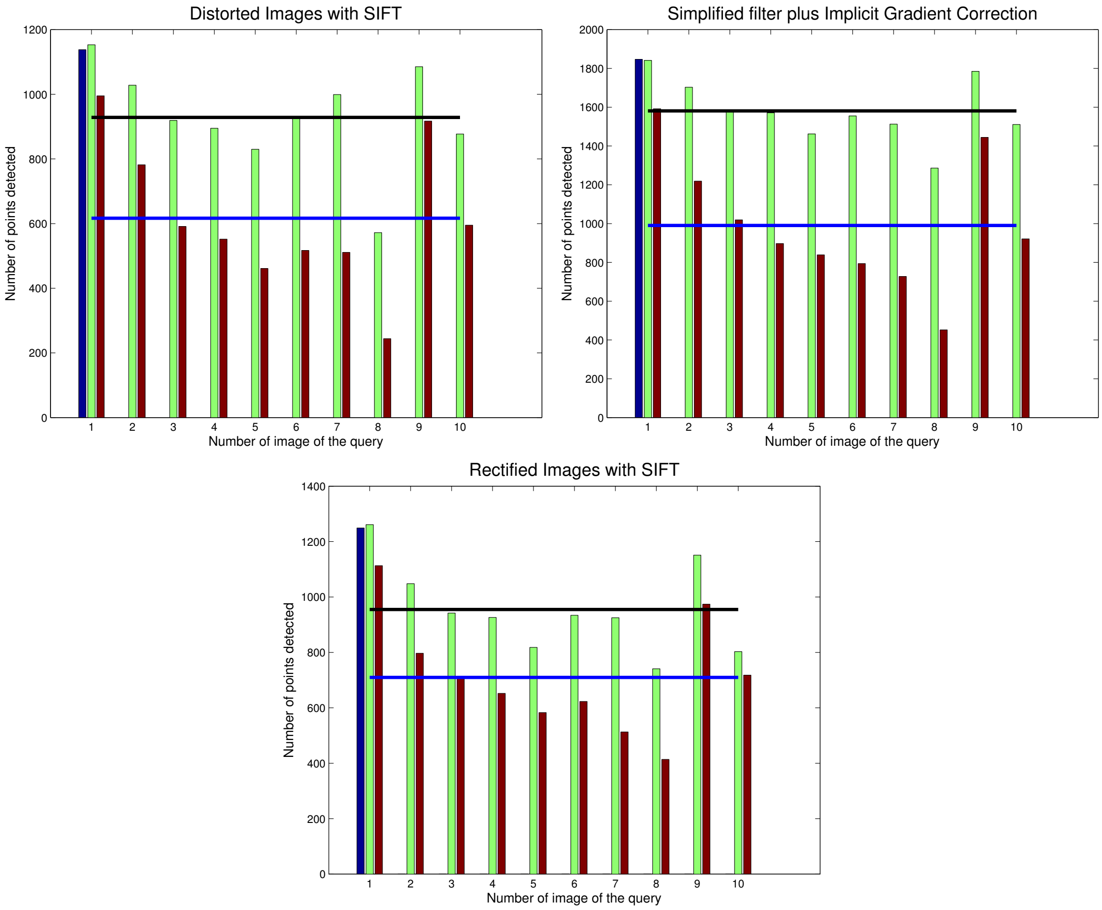


Figure 5.4: Results for detection and matching for three methods in comparison. The blue bar stands for the number of keypoints detected in the Image0 (Fig.(5.4)), green bars for the number of keypoints detected in the query images of Appendix B and the red ones for the maximum number of keypoints matched between the image0 and the correspondent image in the query. The black line averages the total number of detections in the query images the blue cyan line the average number of matches

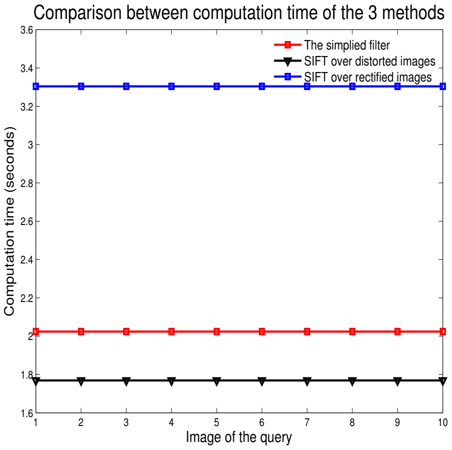
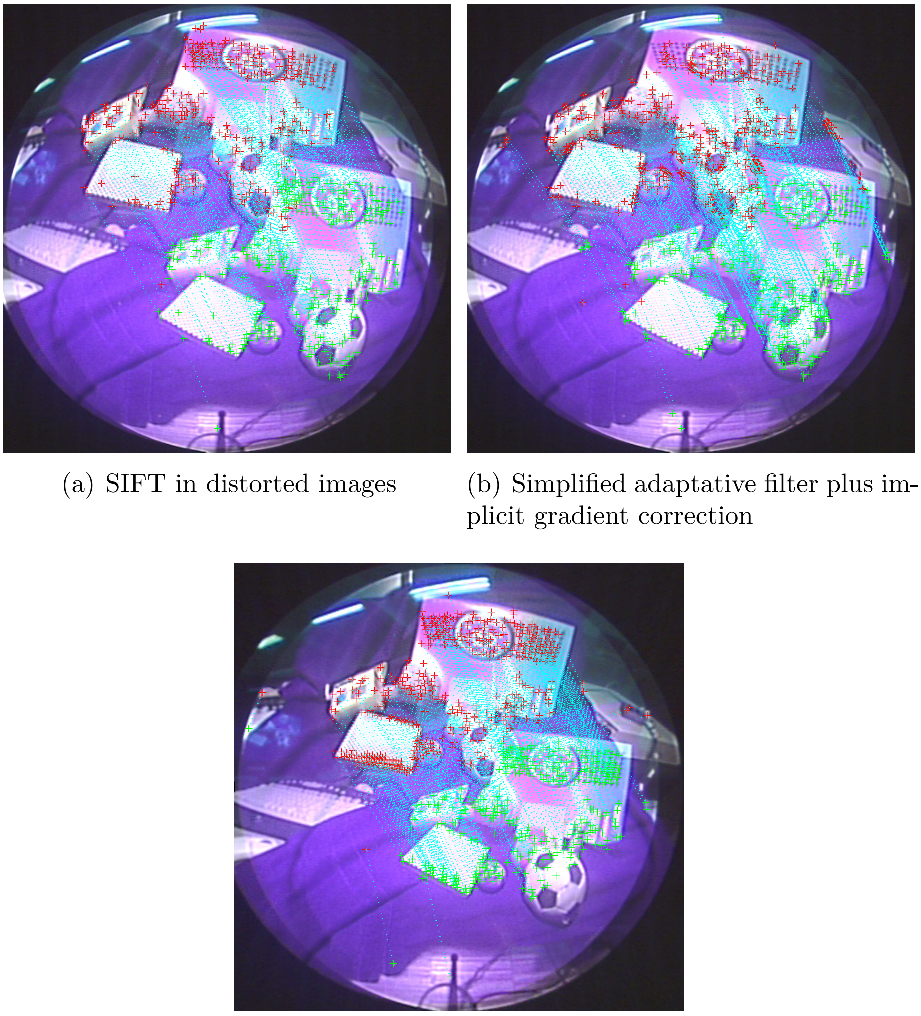


Figure 5.5: The simplified filter performs close the naïve SIFT over radial distortion. It is obvious that with bigger images our method will turn more slow than the SIFT. The variations of the processing time of each image of the query are of milliseconds and could neglected could be neglected comparing with the average time for each method.

the implicit gradient correction under rotation and translation of the arthroscopic lens. Just to show a visual improvement of the proposed method, the image0 will be matched with the an image captured under rotation and translation of the arthroscopic lens. It is observed in Fig.(5.6) that the standard approaches allow to have more matches in the central part of the image. The proposed approach allows to have more points matches in the periphery of the image. However, it must be regarded that the SIFT method doubles the size of the image prior to keypoints selection which contributes with larger structures in the image. The proposed method do not require any interpolation before smoothing the image. The structures of the image are kept at the same size as the original image and the interpolation is only needed to reduce the number of new detections (avoid compressive effect in pixels induced by RD). Our method starts filtering at the same scale as SIFT. However, it is possible to simulate the doubling of the image size. For this, the initial value of sigma could be reduce to the interval $\sigma_0 \in [0.8, 1.2]$, without lost of performance for the descriptor compensation derived.

From the experimental results (simulation and real cases), it can conclude that the method is a suitable approach for use in cameras where the lens induce radial distortion. The detection and descriptor correction methods always outperforms applying SIFT in radial distorted images and for lower levels of distortion also the rectification from distortion. Our method is a computation efficient approach that allows better results for moderate amounts of radial distortion.



(a) SIFT in distorted images

(b) Simplified adaptative filter plus implicit gradient correction

(c) Rectification of RD images

Figure 5.6: Matching between two arthroscopic imagesof the arthroscopic lens. The proposed methods inliers was estimated in 715, beating the two standard approaches with 615 and 495, for the rectification and SIFT over RD images, respectively.

Chapter 6

Conclusions and Outlook

The final derivable of this thesis is an algorithm that partial copes the problem of radial distortion concerning the detection and matching between scenes. The new algorithm is capable of outperform the two classic approaches for lower levels of distortion. When the algorithm was built, all the basic theory of SIFT was used, so it is more than possible to achieve better results and applying the proposed method.

It was also presented an evaluation of SIFT, under non-linear geometric deformation, focus on the case of radial distortion. This is a novel study that was not presented before and it is one of the main contribution of this work. The major problems for SIFT detection and description under RD images are the following:

- The repeatability of the detection and the detections at wrong scales are due to the compressive effect introduced by radial distortion. While the loss of repeatability is more pronounced at lower levels of the DoG pyramid, the wrong scale detections are more noticeable at coarse levels of scale.

- The compression induced by RD in the image spectrum leads to unstable detections that do not occur on the original image.

- The descriptor lose in performance also due to the shifting according to the center of distortion and because the window of the Gaussian weighting is maintained constant all over the image.

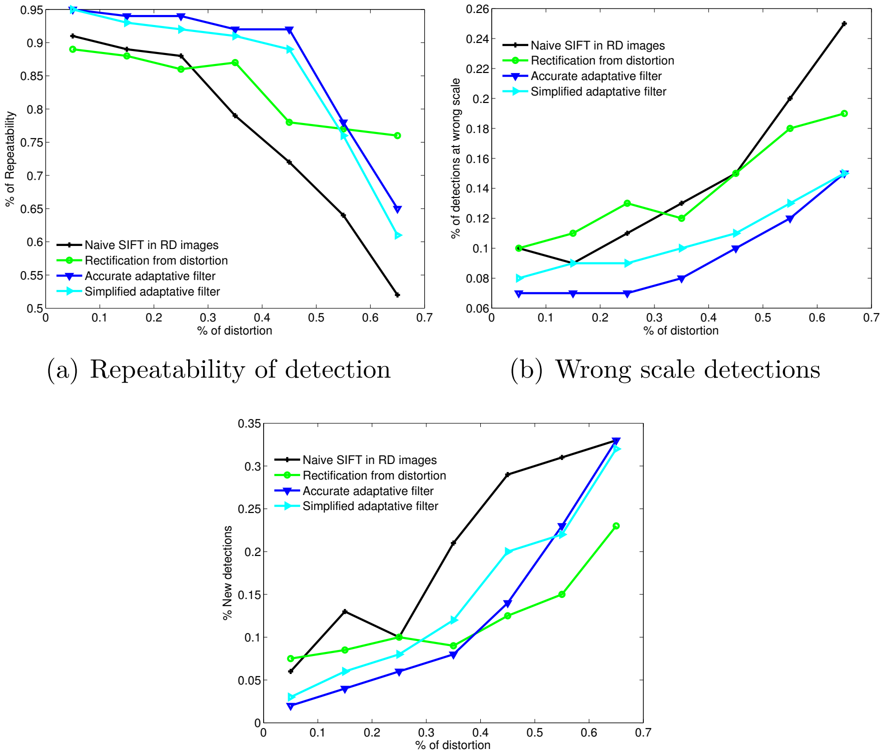
A solution using a Gaussian adaptative filter for detection was derived and a experimental evaluation of its performance was presented. This is the first work that propose to extend SIFT to RD images with all the processing being done in the image plane. However, this approach presents as main drawback the fact of distortion must roughly known. In future steps, the study of the keypoints orientation to find a pattern in their behavior under RD would help to discard the *a priori* information of the RD coefficient.

The gradient correction for the SIFT descriptor computation presents also a suitable approach. It performs better than image rectification and it is a suitable approach to integrate with the detectors presented. However, for higher amount of distortion, the implicit correction starts to present the same effect has the original descriptor computed over the distortion image. One possible solution for higher values of distortion is the projection the corrected gradients to an undistorted space. This step will not require gradient interpolation, and could be performed by the inverse distortion mapping just to avoid the compressive effect over the gradients of the image. This idea was explored by Daniilidis for his SIFT compensation [29] and also by Mikolajczyk [1] to his affine descriptor. This idea can also be extended to our case.

Appendix A

Experimental Results for data set 2

In this appendix are shown the remaining results for the experimental evaluation of the data set 2.

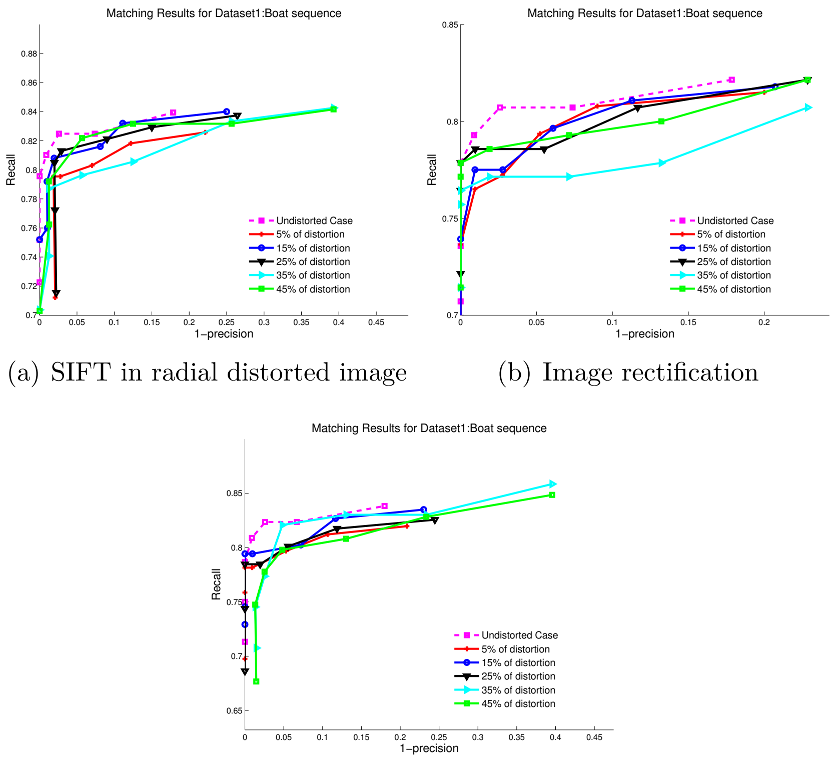


(a) Repeatability of detection

(b) Wrong scale detections

(c) New detections

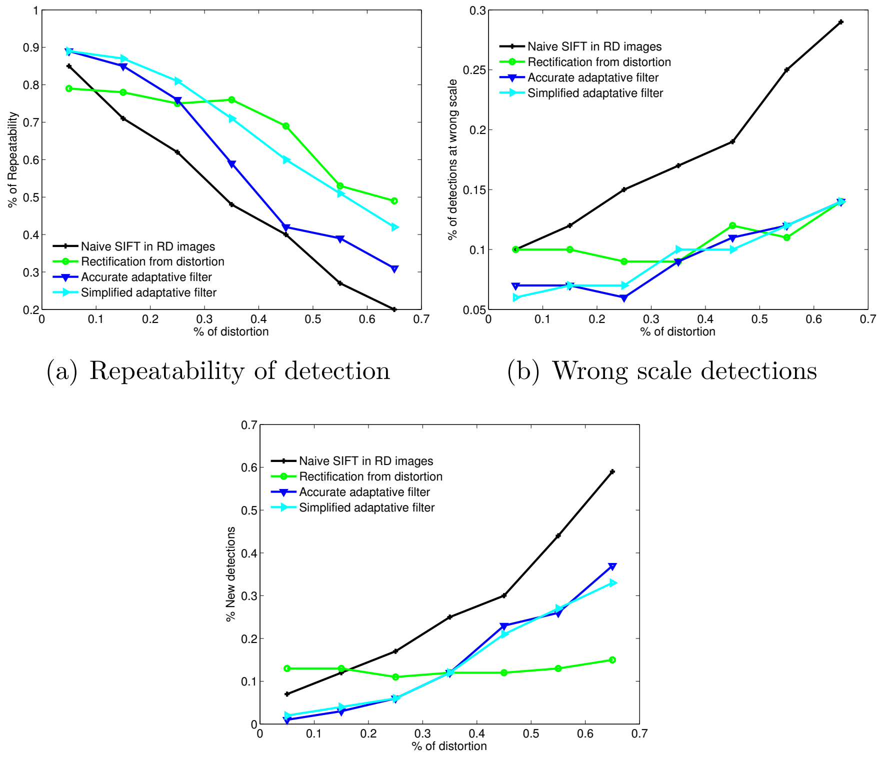
Figure A.1: Detection evaluate for boat1-boat3 case. Our method performs better than the two classic approaches for considerable amounts of distortion.



(a) SIFT in radial distorted image (b) Image rectification

(c) Implicit gradient correction

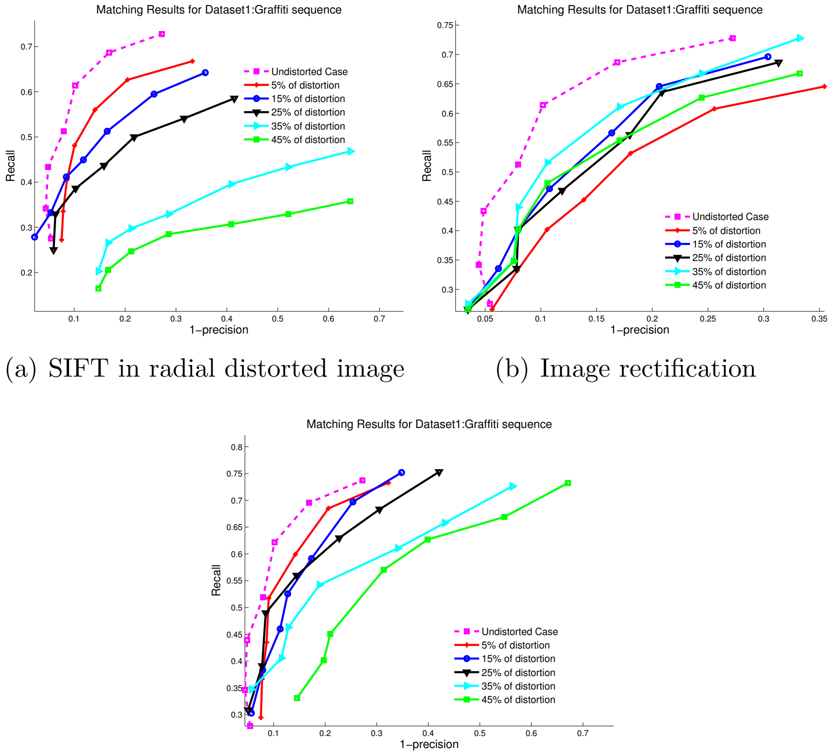
Figure A.2: Recall curves for boat1-boat3 case. In this case, where the distortion is not that meaningful in terms of descriptor computation, our method show similiar performance to naïve SIFT descriptor.



(a) Repeatability of detection (b) Wrong scale detections

(c) New detections

Figure A.3: Detection evaluate for graffiti1-graffiti3case. Once more, the proposed method performs better in terms of repeatability scores for moderate levels of distortion.



(a) SIFT in radial distorted image

(b) Image rectification

(c) Implicit gradient correction

Figure A.4: Recall curves for graffiti1-graffiti3 case. The method in (c) is the most suitable approach at lower levels of distortion (≈ 25% of distortion).

Appendix B

Arthroscopic images data set

This appendix show the arthroscopic images used for the real case experiments.

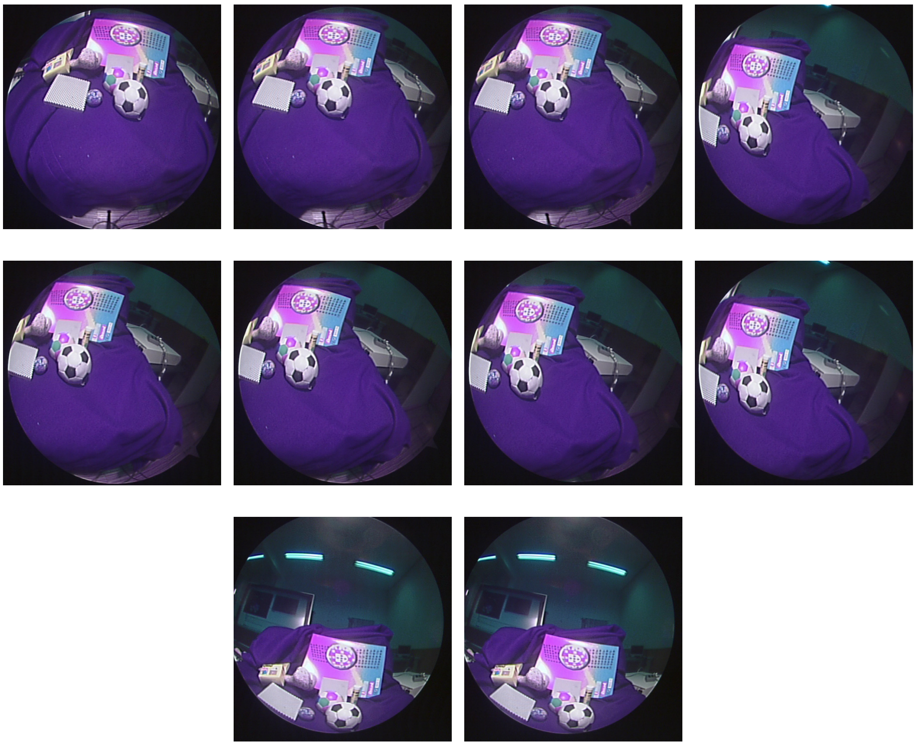


Figure B.1: Arthroscopic images dataset.

Bibliography

[1] K. Mikolajczyk, “Detection of local features invariant to affine transformations,” Ph.D. dissertation, INPG, Grenoble, July 2002. [Online]. Available: http://perception.inrialpes.fr/Publications/2002/Mik02

[2] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] J. P. Barreto and H. Araujo, “Geometric properties of central catadioptric line images and their application in calibration,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1327–1333, 2005.

[4] J. Canny, “A computational approach to edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[5] C. Harris and M. Stephens, “A combined corner and edge detection,” in *Proceedings of The Fourth Alvey Vision Conference*, 1988, pp. 147–151. [Online]. Available: http://www.csse.uwa.edu.au/~{}pk/research/matlabfns/Spatial/Docs/Harris/A_Combined_Corner_and_Edge_Detector.pdf

[6] T. Lindeberg, *Scale-Space Theory in Computer Vision.* Norwell, MA, USA: Kluwer Academic Publishers, 1994.

[7] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. of the International Conference on Computer Vision ICCV, Corfu*, 1999, pp. 1150–1157. [Online]. Available: http://citeseer.ist.psu.edu/lowe99object.html

[8] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *Int. J. Comput. Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[9] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, “A comparison of affine region detectors,” *International Journal of Computer Vision*, vol. 65, no. 1/2, pp. 43–72, 2005. [Online]. Available: http://lear.inrialpes.fr/pubs/2005/MTSZMSKG05

[10] P. Hansen, P. Corke, W. Boles, and K. Daniilidis, “Scale-invariant features on the sphere,” Oct. 2007, pp. 1–8.

[11] “Spherical diffusion for 3d surface smoothing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 12, pp. 1650–1654, 2004, member-Bulow, Thomas.

[12] M. S. Extremal, J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide baseline stereo from,” in *In British Machine Vision Conference*, 2002, pp. 384–393.

[13] Y. Ke and R. Sukthankar, “Pca-sift: a more distinctive representation for local image descriptors,” *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–506–II–513 Vol.2, June-2 July 2004.

[14] E. N. Mortensen, H. Deng, and L. Shapiro, “A sift descriptor with global context,” vol. 1, 2005, pp. 184–190 vol. 1. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1467266

[15] A. Fitzgibbon, “Simultaneous linear estimation of multiple view geometry and lens distortion,” vol. 1, 2001, pp. I–125–I–132 vol.1.

[16] J. P. Barreto, “A unifying geometric representation for central projection systems,” *Comput. Vis. Image Underst.*, vol. 103, no. 3, pp. 208–217, 2006. [Online]. Available: http://www.isr.uc.pt/~jpbar/

[17] D. Burschka, M. Li, R. H. Taylor, and G. D. Hager, “Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery,” in *MICCAI (2)*, 2004, pp. 413–421.

[18] J. P. Barreto and K. Daniilidis, “Fundamental matrix for cameras with radial distortion,” in *ICCV ’05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1.* Washington, DC, USA: IEEE Computer Society, 2005, pp. 625–632.

[19] R. Castle, D. Gawley, G. Klein, and D. Murray, “Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras,” April 2007, pp. 4102–4107.

[20] J. Babaud, A. P. Witkin, M. Baudin, and R. O. Duda, “Uniqueness of the gaussian kernel for scale-space filtering,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 1, pp. 26–33, 1986.

[21] A. P. Witkin, “Scale-space filtering.” in *8th Int. Joint Conf. Artificial Intelligence*, vol. 2, Karlsruhe, August 1983, pp. 1019–1022.

[22] J. J. Koenderink, “The structure of images,” *Biological Cybernetics*, vol. V50, no. 5, pp. 363–370, 1984. [Online]. Available: http://dx.doi.org/10.1007/BF00336961

[23] T. Lindeberg, “Scale-space theory: A basic tool for analysing structures at different scales,” *Journal of Applied Statistics*, vol. 21, pp. 224–270, 1994.

[24] ——, “Feature detection with automatic scale selection,” *Int. J. Comput. Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[25] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005. [Online]. Available: http://lear.inrialpes.fr/pubs/2005/MS05

[26] M. Brown and D. Lowe, “Invariant features from interest point groups,” in *In British Machine Vision Conference*, 2002, pp. 656–665. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.8475

[27] T. Tuytelaars and K. Mikolajczyk, “Local invariant feature detectors: a survey,” *Found. Trends. Comput. Graph. Vis.*, vol. 3, no. 3, pp. 177–280, 2008.

[28] P. Hansen, W. Boles, and P. Corke, “Spherical diffusion for scale-invariant keypoint detection in wide-angle images,” in *DICTA ’08: Proceedings of the 2008 Digital Image Computing: Techniques and Applications.* Washington, DC, USA: IEEE Computer Society, 2008, pp. 525–532.

[29] P. Hansen, P. Corke, W. Boles, and K. Daniilidis, “Scale invariant feature matching with wide angle images,” in *IROS.* IEEE, 2007, pp. 1689–1694.

[30] R. Willson and S. Shaffer, “What is the center of the image,” *Int. Conf. on Computer Vision and Pattern Recognition*, 1993.

[31] 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/>.