UNIVERSITY OF COIMBRA
FACULTY OF SCIENCES AND TECHNOLOGY
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

António Miguel Marques Rodrigues Teixeira Lourenço

KEYPOINT DETECTION, MATCHING, AND TRACKING IN IMAGES WITH NON-LINEAR DISTORTION: APPLICATIONS IN MEDICAL ENDOSCOPY AND PANORAMIC VISION

Tese de Doutoramento em Engenharia Electrotécnica e de Computadores, ramo de especialização em Automação e Robótica, orientada pelo Professor Doutor João Pedro de Almeida Barreto e apresentada ao Departamento de Engenharia Electrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Abril de 2015

UNIVERSIDADE DE COIMBRA

UNIVERSITY OF COIMBRA
FACULTY OF SCIENCES AND TECHNOLOGY
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Keypoint Detection, Matching, and Tracking in Images with Non-linear Distortion: Applications in Medical Endoscopy and Panoramic Vision

Miguel Lourenço

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy

Advisor:
Prof. Dr. João Pedro de Almeida Barreto

April 2015

Acknowledgements

Many people have contributed to make this thesis a reality, by providing me with their guidance, friendship, love, money, code and data. To those people, brave of the bravest, an acknowledgement word shall be given in this very pages.

I would like to start by thanking my advisor Prof. João P. Barreto for teaching me a great deal about science, and how to tackle hard problems in computer vision. His enthusiasm and pressure for having new exciting results was a key ingredient to put this thesis on the right track.

I also acknowledge the Portuguese Governments department FCT- Portuguese Foundation for Science and Technology for the doctoral scholarship SFRH/BD/63118/ 2009 that funded my work. After a four year PhD and a 12 month research grant I admit that I owe a lot to Portuguese tax payers. Currently I am contributing with security at airport electronic gates, but I hope I can keep helping people lives with computer vision systems.

A special word to my co-authors, specially to Prof. Danail Stoyanov for having the patience to read my paper drafts and for providing the data for some of my papers.

I would like to thank my big family for the long standing and tirelessly support. Specially to my awesome nephews for keeping me smiling and asking me for money for their video games or their bike.

Thanks to my lab friends, Melo, Michel, Luis, and Vitor...and many others that somehow made this journey more enjoyable. Specially thanks to my personal friends Claudia, Pedro and Rita for helping my wife in keeping me fed during these years with some awesome dinners and great wine bottles.

The most important *thank you* goes to my wife Margarida, whose unconditional love and support would never fit in these pages. Meeting you at the end of my days during my extreme paper deadlines was always quite comforting. Right now, we are engaged in the most wonderful process of all: building a family. So my truly last words are to my daughter Beatriz, who is giving my life a whole new meaning.

Abstract

Point correspondences between different views are the input to many computer vision algorithms with a multitude of purposes that range from camera calibration to image content retrieval, and pass by structure-from-motion, registration, and mosaicking. Establishing such correspondences is particularly difficult, not only in the case of wide-baseline and/or strong change in viewpoint, but also when images present significant non-linear distortions. The thesis addresses this last problem and investigates solutions for detecting, matching, and tracking points in images acquired by cameras with unconventional optics such as fish-eye lenses, catadioptric sensors, or medical endoscopes.

We start by studying the impact of radial distortion in keypoint detection and description using the well known SIFT algorithm. Such study leads to several modifications to the original method that substantially improve matching performance in images with wide field-of-view. Our work is conclusive in showing that non-linear distortion must be implicitly handled by a suitable design of filters and operators, as opposed to being explicitly corrected via image warping. The benefits of such approach are demonstrated in experiments of structure-from-motion, as well as in the development of a vision-system for indoor localization where perspective images are used to retrieve panoramic views acquired with a catadioptric camera.

In a second line of research, we investigate solutions for feature tracking in continuous sequences acquired by cameras with radial distortion. We build on the top of the conventional frameworks for image region alignment and propose specific deformation models that simultaneously describe the effect of local image motion and global image distortion. It is shown for the first time that image distortion can be calibrated at each frame time instant by tracking a random set of salient points. The result is further explored to solve the problem of knowing the intrinsic calibration of cameras with motorised zoom at all times. This problem is particularly relevant in the context of medical endoscopy and the solution passes by combining off-line calibration with on-line tracking to update of the camera focal length. The effectiveness of our tracking and calibration approaches are validated in both medical and non-medical video sequences.

The last contribution is a pipeline for visual odometry in stereo laparoscopy that relies in multi-model fitting for segmenting different rigid motions and implicitly discarding regions of non-rigid deformation. This is complemented by a temporal clustering scheme that enables to decide which parts of the scene should be used to estimate the camera motion in a reliable manner.

Resumo

Correspondências de pontos entre imagens da mesma cena são o argumento de entrada para muitos algoritmos de visão por computador, como por exemplo calibração de câmaras, reconhecimento de imagens e recuperação de movimento e estrutura 3D da cena. O cálculo de correspondências é particularmente difícil, não só devido a deslocamentos de câmara e mudanças de ponto de vista, mas também devido à presença de deformação não-linear, como é o caso de distorção radial. Esta tese investiga o último problema e propõe soluções para deteção, correspondência e seguimento de pontos em imagens adquiridas com câmaras equipadas com ópticas não convencionais, como lentes olho-de-peixe, sensores catadióptricos e endoscópios/ laparoscópios médicos.

Esta tese começa por estudar o impacto da distorção radial na deteção e descrição de pontos de interesse do método SIFT. Este estudo leva a várias modificações ao método original que permitem melhorias substanciais no desempenho em imagens adquiridas com câmaras com largo campo de visão. É demonstrado que a distorção não-linear deve ser implicitamente compensada atráves da adaptação dos operadores de imagem em vez de rectificar as imagens para a remover. Os benefícios desta nova solução são validados com experiências de recuperação de movimento e através de um sistema de visão que usa uma base de dados de imagens catadióptricas georeferenciadas para reconhecimento de localizações dentro de edifícios.

Numa segunda linha de investigação são estudadas soluções para seguimento de pontos de interesse em sequências contínuas de imagens com distorção radial. Usando como base o actual estado da arte para registo de imagens, são propostas soluções para descrever simultaneamente o efeito do movimento local e distorção global da imagem. É demonstrado pela primeira vez que a distorção radial na imagem pode ser calibrada em cada instante de tempo através do seguimento de pontos de interesse. Esta solução é ainda explorada para resolver o problema de calibração de câmaras com zoom motorizado. Este problema é particularmente relevante no contexto de endoscopia médica e a solução passa por combinar calibração *offline* com calibração *online* usando o seguimento de pontos para actualizar a distância focal da câmara. A eficácia dos algoritmos de seguimento e calibração são validados em sequências de vídeo médicas e não-médicas.

A última contribuição desta dissertação é um método para odometria visual em laparos- copia éstereo que utiliza técnicas de estimação de mútiplos modelos para segmentar a cena em zonas rígidas e não-rígidas. De modo a complementar a segmentação inicial um esquema de *clustering* temporal é usado para decidir quais zonas da cena devem ser utilizadas para ancorar a estimação do movimento da câmara.

Contents

List of Figures

List of Tables

List of Acronyms

RD Radial Distortion

SfM Structure-from-Motion

FOV Field-of-View

BOV Bags of Visual Words

SIFT Scale-Invariant Feature Transform

sSIFT Spherical SIFT

pSIFT Approximated Diffusion SIFT

LB Laplace-Beltrami

RD-SIFT Radial Distortion and Scale Invariant Feature Transform

sRD-SIFT separable RD-SIFT

RANSAC RANdom SAmple Consensus

GVP Geometric Preserving Visual Phrases

cylSIFT SIFT for Cylindrical Images

VCP Virtual Camera Perspectives

KLT Lucas-Kanade-Tommasi tracker

cRD-KLT calibrated Radial Distortion KLT

uRD-KLT uncalibrated Radial Distortion KLT

MIS Minimally Invasive Surgery

CAS Computer Assisted Surgery

Chapter 1

Introduction

Images are low dimensional representations of the 3D world. The extraction of visual content from these low dimensional representations enable to design algorithms for interpreting the observed scene. One way of extracting image information is by computing image salient point, commonly called interest points or keypoints, that can be easily recognized across different views of the same scene. The literature reports a large pool of approaches for finding image salient points that can range from low computationally complexity translation invariants [1,2] to more complex solution that achieve scale [3, 4] or affine invariance [5, 6]. Such features capture low-level local image information that can be used for a wide range of computer vision applications such as image retrieval [7,8], classification, point association for structure-from-motion [9–12], or image compression [13, 14].

In the context of medical endoscopy, image features have been used mainly with two purposes: content retrieval, and stucture-from-motion and navigation. Image retrieval and classification of different medical image modalities using computer vision techniques has been proposed in the literature [15–17]. Visual assessment of medical images performed by physicians is often subjective and experience dependent. An image retrieval system with an annotated database can be used by inexperienced medical professionals as an auxiliary tool for disambiguating difficult diagnosis. The importance of medical image retrieval has been recently acknowledged with the creation of ImageCLEF [18] context. This benchmark provides several medical image modali-

(a) Fish-eye lens (b) Micro lens (c) Endoscopic lens

Figure 1.1: Different types of lenses that induce strong RD.

ties for evaluation of retrieval schemes, and it was created for a fast dissemination of medical image retrieval applications.

Knowledge about the camera motion is a fundamental prerequisite for image-based computer aided-surgery, enabling 3D reconstruction [19], registration with other sensor modalities [20], and guidance systems. Typical approaches compute point association between views and uses the 5-point algorithm for solving the camera motion within a sample consensus framework [19–21].

Endoscopic images place several challenging problems for keypoint detection and matching that are not currently solved by state-of-the-art techniques. One of such problems is the strong non-linear distortions, namely Radial Distortion (RD), resulting from the miniaturization and special optical arrangement of endoscopic lens. In [19, 20] the authors deal with this problem by previously correcting the images to remove the radial distortion introduced by the medical endoscopes/laparoscopes. In this dissertation we show that the image rectification must be avoided since it has a negative impact in terms of feature accuracy. Thus, we propose computational efficient approaches for keypoint detection, matching, and tracking in images with RD that significantly improve the accuracy of Structure-from-Motion (SfM) application in medical images.

Despite of the fact that we are primarily motivated by medical endoscopic applications, the research we will pursue has a more general character. Many vision systems employ cameras with unconventional optical arrangements that introduce non-linear distortions. The most striking example is the case of cameras equipped with fish-

eye lenses (see Fig. 1.1(a)) for the acquisition of wide Field-of-View (FOV) images that enable a more thorough visual coverage of the environments. Another example of unconventional optics is the case of cameras equipped with micro-lenses (see Fig. 1.1(b)) and boroscopes for visual inspection of cavities with limited access. The evaluation in non-medical scenarios using standard evaluation guidelines is carried to prove that the usefulness of the proposed solutions goes beyond its application the medical endoscopic images.

1.1 Thesis Outline and Contributions

This document presents three main lines of research.

In the first line, we focus on keypoint detection and matching in sparse image sequences. **Chapter 2** starts by benchmarking the performance of the Scale-Invariant Feature Transform (SIFT) in images with distortion. Afterwards, we propose improvements to the SIFT algorithm that substantially improve keypoint detection and matching in images with RD. Since, the design of such solution is based in adapting the low-level image processing tools to account for the distortion effect, the benefits are transversal to all methods using scale-space image representation and gradient-based keypoint descriptors. This chapter is closely related with the following conference and journal publications:

- M. Lourenco, J.P. Barreto, and A. Malti, *Feature Detection and Matching in Images with Radial Distortion*, IEEE International Conference on Robotics and Automation, 2010. (Finalist for the Best Student Paper Award).

- M. Lourenco, J.P. Barreto, and F. Vasconcelos, *sRD-SIFT Keypoint Detection and Matching in Images with Radial Distortion*, IEEE Transactions on Robotics, 2012.

Chapter 3 extends the methodology proposed in the previous chapter to para-catadioptric images, and proves its usefulness in a indoor image-based localization application. We propose to carry the localization by using perspective images acquired with a cell phone to query a database of geo-referenced panoramic images.

The retrieval is accomplished by using a combination of a new keypoint detection and matching strategy with a state-of-the-art Bags of Visual Words (BOV) recognition engine. This chapter relates with the following conference publication:

- M. Lourenco, V. Pedro, and J.P. Barreto, *Localization in Indoor Environments by Querying Omnidirectional Visual Maps using Perspective Images*, IEEE International Conference on Robotics and Automation, 2012.

The second line of research is devoted to feature association in continuous endoscopic video. In **Chapter 4** we study the problem of image alignment in the presence of non-linear distortion. We build on top of the conventional image registration frameworks, and propose specific deformation models that simultaneously describe the effect of local image motion and global image radial distortion. We propose solutions for both calibrated and uncalibrated camera setups, showing that it is possible to reliably estimate the image distortion at each frame time instant by tracking random keypoints. The proposed methods are evaluated in a feature tracking context showing to highly benefit standard rigid SfM in medical endoscopy. In **Chapter 5** we have extended the uncalibrated image alignment framework for estimating the focal length in cameras equipped with optical zoom, showing that is possible to accurately keep the camera calibrated by combining off-line camera calibration with on-line image distortion estimation. These two chapters are related with two conference publications and a patent request:

- M. Lourenco and J.P. Barreto, *Tracking Feature Points in Uncalibrated Images with Radial Distortion*, European Conference on Computer Vision, 2012.

- M. Lourenco *et al.* , *Continuous Zoom Calibration by Tracking Salient Points in Endoscopic Video*, International Conference on Medical Image Computing and Computer Assisted Intervention, 2014 (Oral presentation).

- M. Lourenco, J.P. Barreto and R. Melo, *Method for aligning and tracking point regions in images with radial distortion that outputs motion model parameters, distortion calibration, and variation in zoom*, PCT/PT2013/000057, 2013.

Finally, **Chapter 6** presents an effective visual odometry solution for stereo laparoscopes. The proposed pipeline effectively segments non-rigid and piecewise rigid structures from the surgical site by using a multi-model fitting [22]. This is complemented by a temporal clustering scheme to better distinguish which scene regions should be used to anchor the camera motion estimation. This work has been published in:

- M. Lourenco, D. Stoyanov and J.P. Barreto, *Visual Odometry in Stereo Endoscopy by using PEaRL to handle Partial Scene Deformation*, International Workshop in Augmented Environments for Computer-Assisted Interventions, held in conjuction with Internation Conference on Medical Image Computing and Computer Assisted Intervention, 2014.

Notation

Matrices are represented by symbols in sans serif font, e.g. M, and image signals are denoted by symbols in typewriter font, e.g. $\mathtt{I}$. Vectors and vector functions are typically represented by bold symbols, and scalars are indicated by plain letters, e.g $\mathbf{x} = (x, y)^\mathsf{T}$ and $\mathbf{f}(\mathbf{x}) = (f_x(\mathbf{x}), f_y(\mathbf{x}))^\mathsf{T}$. We use under script, i.e. $\mathbf{f}_\mathbf{p}(\mathbf{x})$, to denote when a certain function parameter $\mathbf{p}$ is known, and $\mathbf{f}(\mathbf{x}; \mathbf{p})$ to denote that $\mathbf{p}$ is unknown. $\mathbf{0}$ is specifically used to represent a null vector.

Chapter 2

Radial Distortion and Scale Invariant Features

Keypoint detection and matching is of fundamental importance for many applications in vision-based systems. The association of points across different views is problematic because image features can undergo significant changes in appearance. Unfortunately, state-of-the-art methods, like the SIFT, are not resilient to the radial distortion that often arises in images acquired by cameras with micro-lenses or wide FOV. This chapter proposes modifications to the SIFT algorithm that substantially improve the repeatability of detection and effectiveness of matching under radial distortion, while preserving the original invariance to scale and rotation.

2.1 Introduction

Finding point correspondences between two images of the same scene is a key step of many computer vision algorithms. Camera calibration, image registration, structure-from-motion, visual recognition, and image content retrieval are just a few examples of applications that use discrete point matches as input. Current methods for associating points across different views typically comprise three steps: (i) the *detection* of keypoints at distinctive locations in the image, such as corners, blobs, and T-junctions.

The most valuable property of a keypoint detector is its ability of repeatedly find the same physical point under different viewing conditions; (ii) the *description* of the neighborhood patch around the detected keypoints. The neighborhood is usually represented through a feature vector that must be distinctive and, at the same time, robust to geometric and photometric transformations; and finally (iii) the *matching* of descriptor vectors which is typically carried using a distance defined in the feature space, e.g. Mahalanobis or Euclidean distance.

The SIFT, introduced by Lowe [3, 23], have become arguably one of the most popular matching algorithms, being broadly used in robotics for tasks like visual servoing and visual Simultaneous Location and Mapping [11, 24], content image retrieval [16] and medical endoscopy structure-from-motion [19, 20]. The detection is carried in a scale-space representation of the image [25] that is efficiently computed using the Difference-of-Gaussian (DoG) operator. The keypoint detection is performed by searching for points in the DoG pyramid that are simultaneously extrema in space and scale dimensions. This procedure enables assigning scale information to salient points, which is used for normalizing the size of the neighborhood region considered during description. The descriptor vector encodes the local image gradients that are expressed with respect to the dominant gradient orientation. The SIFT features obtained in this manner are invariant to scale, rotation, illumination, and moderate viewpoint changes.

Many vision-based systems employ cameras with unconventional optical arrangements that introduce non-linear distortions. The most striking example is the case of cameras equipped with fish-eye lenses for the acquisition of wide FOV images. Such cameras enable a thorough visual coverage of the environments, and are advantageous for egomotion estimation by avoiding the ambiguity between translation and rotation [26, 27]. Another example of unconventional optics is the case of cameras equipped with micro-lenses and boroscopes for visual inspection of cavities with difficult or limited access [28]. These cameras are broadly used in medicine for endoscopic procedures of surgery and diagnosis. Unfortunately the SIFT algorithm, as well as the majority of competing methods, is meant for perspective images and cannot handle the strong radial distortion introduced by the optics described above [29–32].

2.1.1 Related Work

The RD is a non-linear geometric deformation caused by the bending of the light rays when crossing the optics. At an image level, and comparing with the standard perspective, the pixel positions suffer a displacement along radial directions and towards the center. This displacement is non-uniform and depends on the distance to the image center (the radius). Despite the fact that the SIFT algorithm is not invariant to RD, it has been applied in the past to images with significant distortion. While ones simply ignore the pernicious effects of RD and directly apply the original SIFT algorithm over distorted images [19], others perform a preliminary correction of distortion through image rectification and then apply SIFT [33]. The latter approach is quite straightforward but it has two major drawbacks: the explicit distortion correction can be computationally expensive for the case of large frames and, more importantly, the interpolation required by the image rectification introduces artifacts that affect the detection repeatability.

Daniilidis *et al.* were the first ones arguing that the warping of wide FOV images should be avoided because interpolation effects introduce undesired results in filtering [29]. Their article proposes using the sphere as the underlying domain of the image function for computing optical flow in catadiotpric views. However, instead of back-projecting the image plane $\mathbb{P}^2$ into the sphere $\mathbb{S}^2$, the smoothing is formulated in $\mathbb{S}^2$ and the derived kernel function is projected into $\mathbb{P}^2$. This enables carrying the convolution on the plane using the original image pixel values. Since the mapped spherical kernel changes at each image pixel position, the computational complexity of the filtering is substantially higher when compared with the convolution with standard isotropic Gaussian kernel, which can be performed separately in X and Y dimensions [34].

In [35, 36] Bulow proposes a scale-space representation for functions defined in $\mathbb{S}^2$ by solving the spherical heat diffusion equation. Inspired by this work, Hansen *et al.* investigated the generalization of the SIFT algorithm for images with domain on the sphere [31, 32]. The advantages of such generalization are twofold: First, the SIFT on the sphere can be indistinguishably applied to any type of central projection image. The only requirement is to know in advance the intrinsic camera calibration in order to map the image plane into $\mathbb{S}^2$; Second, the formulation of SIFT on the sphere

enables to achieve full invariance to pure camera rotation motion. The original SIFT algorithm [3], despite of being invariant to rotations on the plane, it is unable to handle the projective transformations in $\mathbb{P}^2$ due to camera rotation [37].

The main difficulty in extending the SIFT algorithm to the sphere is the computation of a suitable scale-space representation that passes by back-projecting the image $\mathtt{I}$ into $\mathbb{S}^2$ and convolving the result with a spherical Gaussian function G_S [36]. The problem is that this operation must be carried in a manner that is simultaneously computationally efficient and avoids the re-sampling of the original image signal [29, 31]. We briefly review the approaches described in the literature:

- *Mapping G_S into $\mathbb{P}^2$:* The re-sampling can be avoided by mapping G_S into $\mathbb{P}^2$ and carrying the convolution on the image plane using the original pixel values. This is similar to the adaptive filtering proposed in [29], with the mapped Gaussian kernel changing at every image pixel location, which precludes the separability property in X and Y. Unfortunately this solution is unsuitable for generating the multiple scale levels of the $\mathtt{DoG}$ pyramid because of its computational complexity [31, 32].

- *Spherical SIFT (sSIFT):* An alternative is to perform the Gaussian smoothing in the frequency domain [31]. Since the original image $\mathtt{I}$ can be mapped into a spherical image $\mathtt{I}_S$, then the spectrum of $\mathtt{I}_S$ can be found via a discrete spherical Fourier transform (DSFT). This means that the filtering can be carried by applying the inverse DSFT to the product of the image spectrum with the transform of G_S. The spherical diffusion can be implemented in an efficient manner in the spectral domain as long as it is imposed an upper limit on the bandwidth for keeping the computation tractable. The problem is that this limit can lead to aliasing issues when finding the image spectrum as discussed in [31, 32].

- *Approximated Diffusion SIFT (pSIFT):*: Hansen *et al.* have recently used stereographic projections for approximating the diffusion on the sphere [32]. They propose to map the image $\mathtt{I}$ via the sphere into the stereographic plane, and convolve the result with the stereographic projection of G_S. The projected Gaussian kernel, despite of changing at every image pixel position, it is always a

symmetric function. More importantly, it is shown that the 2D adaptive filtering is well approximated by successive 1D convolutions along X and Y directions. This enables to achieve a computational efficiency similar to the original SIFT, while avoiding the aliasing problems of the spectral approach. Although not discussed in [32], the method has the drawback of requiring image re-sampling for mapping `I` into the stereographic plane.

- *Laplace-Beltrami (LB)* operator [38–40]: Recently, some authors applied Riemannian geometry concepts for computing the scale-space representation for images of central projection systems. The Gaussian smoothing on the sphere is achieved through an iterative procedure that preserves the geometry of the visual contents and adapts to the non-uniform resolution while using the original image pixel values. Although the Laplace-Beltrami can be derived for any central projection system, the iterative procedure required to smooth the image signal is computational expensive as shown in [40].

The Radial Distortion and Scale Invariant Feature Transform (RD-SIFT) herein presented consists on several well engineered modifications to the original SIFT framework for improving its invariance to radial distortion. Every processing step is carried on the plane using original pixel values and, in a similar manner to the pSIFT algorithm, the computational efficiency of the adaptive filtering is improved by considering an approximate kernel function that is separable in X and Y directions. Comparing with the SIFT formulations on the sphere, the RD-SIFT is less general, in the sense that it cannot be applied to images not following the division model (e.g. catadioptrics), and is not invariant to the effects of pure camera rotation motion. However, and unlike the sSIFT and pSIFT algorithms, the RD-SIFT neither has bandwidth limitations, nor requires warping of the original image. This difference seems to play a key role in terms of matching performance. Hansen *et al.* compared sSIFT and pSIFT against the original SIFT in sequences acquired by a fish-eye camera, and reported improvements in matching performance of at most 15% [32]. As shown in section 2.6, the RD-SIFT algorithm can improve the effective number of matches up to 50% when compared with the original SIFT algorithm. Another advantage of RD-SIFT with respect to SIFT formulations on the sphere is that it does not require accurate

intrinsic camera calibration (an approximate modelling of the distortion is sufficient).

2.1.2 Chapter Overview

The next sections start by introducing the camera projection model adopted along this document, and briefly reviews the original SIFT algorithm [3, 23].

Section 2.3 evaluates and discusses the effects of RD in SIFT detection and description. The experiments are carried on a representative set of perspective images to which distortion is artificially added. The usage of synthetically distorted images enables fully controlled experiments, with accurate ground truth and assurance that the observations are only due to the influence of RD. It is shown that SIFT detection is affected in multiple manners, with some keypoints, previously found at fine scales, being missed; others being assigned to incorrect scales; and false keypoints being detected because of spurious image artifacts due to the distortion (e.g. straight lines that become curves). Most of these observations can be qualitatively explained by the non-uniform compression of image structures. The compression diminishes the characteristic length of the features and, as a consequence, the extrema in the DoG pyramid tend to occur at scales that are lower than they would be in the absence of distortion. In addition, and since RD also modifies the image gradients, the vector description varies with the position where the feature is projected. Therefore, it is easy to understand that distortion also affects negatively the SIFT matching performance.

Sections 2.4 and 2.5 suggest modifications to the SIFT framework that substantially improve its resilience to non-linear distortion. Section 2.4 focus on the detection, while section 2.5 concerns feature description and matching. A straightforward solution for the RD problem consists in correcting the distortion followed by carrying the keypoint detection and description in the rectified image. However, the explicit distortion correction requires image re-sampling and, as discussed in [29, 32], the pixel interpolation affects the DoG filtering output, which influences the repeatability of keypoint detection. We propose instead to filter the original frame by an adaptive kernel that takes into account the RD at each image pixel position. This approach outperforms the explicit distortion correction because it avoids the signal reconstruction. It is also shown that the adaptive filtering can be well approximated by horizontal

and vertical 1-D correlation using a Gaussian kernel with standard deviation varying with the pixel image radius. Such approximation enables a computational efficiency that is comparable to the original SIFT algorithm. Following a similar philosophy, section 2.5 proposes to achieve description invariance to RD by performing implicit gradient correction using the Jacobian of the distortion function. Finally, section 2.6 conducts several tests using real distorted images that prove the superiority and usefulness of RD-SIFT, and shows that the proposed modifications preserve the original SIFT invariance to scale, rotation, illumination, and small viewpoint changes.

2.2 Background

2.2.1 The Scale Invariant Feature Transform

The SIFT framework was originally introduced by Lowe in [3]. The keypoint detection is carried in a scale-space representation of the image [25], which enables associating scale information to points that are visually salient. The scale is used for normalization purposes during the description stage. The descriptor of each keypoint is a vector that encodes the image gradients on a local patch around the point. The size of this patch depends on the scale of selection (invariance to scale), and the local gradients are described with respect to a dominant gradient orientation (invariance to rotation). The SIFT detection and description are further discussed below:

SIFT detector

SIFT relies in the scale-space theory for achieving scale invariance and high repeatability in detection [25]. Lowe [3] uses the DoG operator for extrema detection, an approximation of the Laplacian-of-Gaussian(LoG) that enable to improve the computational efficiency and avoid the explicit computation of second order derivatives that are highly sensitive to noise [34]. Let $\mathtt{L}$ be a blurred version of $\mathtt{I}$ obtained by convolution with a 2D Gaussian function with standard deviation σ (the scale).

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

Each level of the DoG pyramid is computed through the subtraction of successive blurred versions of I.

$$\mathtt{DoG}_{k^{n+1}\sigma}(x,y) = \mathtt{L}_{k^{n+1}\sigma}(x,y) - \mathtt{L}_{k^{n}\sigma}(x,y)\,, \tag{2.2}$$

where k denotes a constant multiplicative factor.

The keypoints are detected by looking for extrema in the scale-space representation of the image signal. The intuition is that an extrema along the space dimension reveals the location of a visual salience. In addition, an extrema along the scale dimension is illustrative of the correlation between the characteristic length of the image feature and the standard deviation of the Gaussian filter. It can be shown that a keypoint with characteristic length $\sqrt{\sigma}$ gives raise to an extrema at the scale level σ [41].

In the SIFT algorithm the search for extrema in the DoG pyramid is performed by comparing each point with its $3 \times 3 \times 3$ neighborhood. Lowe [3] suggests to double sample the initial image I in order to increase the number of extrema detections. This corresponds to a scaling of the spectrum in the frequency domain, which enables the capture of high frequency component by the DoG band-pass filtering. Unfortunately some of these extrema are either artifacts due to the bilinear interpolation, or lie in indistinguishable image regions (e.g. low contrast regions or non discriminant edges). These extrema are discarded [42], and the position of the detected keypoints is refined to sub-pixel precision through interpolation in the DoG domain.

SIFT descriptor

The detection stage provides the image coordinates $\mathbf{x}$ and scale σ of a set of keypoints. The following step is assigning to each keypoint a descriptor vector that encodes the image gradients on a local patch around the point. The size of the patch is defined by the scale of selection σ, and the entire processing is carried at the level of the Gaussian pyramid where the extrema occurred (scale invariance). The window is divided in a 16×16 neighborhood and the gradient magnitude and orientation is computed at each point of the grid.

The description algorithm starts by determining the dominant orientation of the

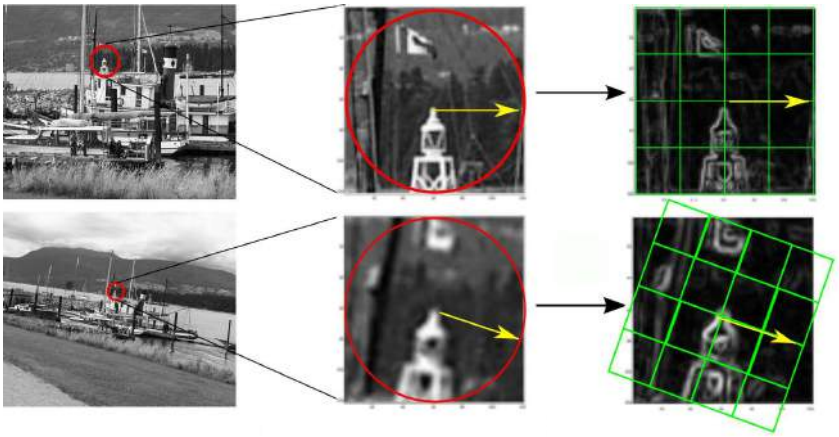


Figure 2.1: Scale and rotation invariant descriptor computation. The extracted scale of the keypoint is used to compute the descriptor, after a rotation normalization step.

gradients, that is used as reference for rotating the window towards a normalized position (rotation invariance). The estimation of the dominant orientation is carried by looking for peaks in a histograms of 36 bins. Each bin represents an interval of 10° around the keypoint $\mathbf{x}$, and accumulates the magnitudes of the gradients whose orientations fall within its range. The gradient samples are weighted by a Gaussian with center $\mathbf{x}$ and standard deviation 1.5σ that aims giving less emphasis to contributions far away from the keypoint. If there is a secondary peak, then a new descriptor is created with the same scale-space information but different orientation. This means that the same keypoint can have more than one associated descriptor which proves to be helpful in improving the robustness during matching .

After compensating for the rotation, the 16×16 neighborhood is divided into 16 sub-regions with size 4×4. Each sub-region gives raise to an histogram of gradient magnitudes where the gradient orientations are quantized into 8 intervals. The final descriptor is obtained by stacking the 16 histograms with 8 bins into a vector with dimension 128. The division into sub-regions enables the descriptor to be invariant to pixel shifts up to 4 positions in the image. The gradient samples are weighted by a Gaussian function with center $\mathbf{x}$ and standard deviation 0.5σ. This prevents sudden changes in the descriptor caused by small changes in the window position, and avoids mutual interferences between keypoints that are spatially close. The filtering

procedure is of key importance for assuring stability and distinctiveness of the final 128-dimensional vector [3].

2.2.2 The Division Model for Radial Distortion

Along this document it is assumed that the image distortion follows the first order division model [43, 44], with the amount of distortion being quantified by a single parameter η (typically $\eta < 0$), and the distortion center being approximated by the image center. Let $\mathbf{x} = (x, y)^\mathsf{T}$ and $\mathbf{u} = (u, v)^\mathsf{T}$ be the coordinates of corresponding points in the distorted and undistorted images expressed with respect to a reference frame with origin in the center. $\mathbf{\Gamma}$ is a vector function that maps distorted points in the distorted image plane $\mathtt{I}$ into points in the undistorted image $\mathtt{I}^u$ [43,44]:

$$\mathbf{u} = \mathbf{\Gamma}(\mathbf{x}) = \Big(1 + \eta\,\mathbf{x}^\mathsf{T}\,\mathbf{x}\Big)^{-1}\,\mathbf{x} \tag{2.3}$$

$$= \begin{pmatrix}\Gamma_u(\mathbf{x})\\ \Gamma_v(\mathbf{x})\end{pmatrix} = \begin{pmatrix}\frac{x}{1+\eta\,(x^2+y^2)}\\ \frac{y}{1+\eta\,(x^2+y^2)}\end{pmatrix}. \tag{2.4}$$

The function is bijective and the inverse mapping from $\mathtt{I}^u$ to $\mathtt{I}$ is given by

$$\mathbf{x} = \mathbf{\Gamma}^{-1}(\mathbf{u}) = 2\Big(1 + \sqrt{1 - 4\,\eta\mathbf{u}^\mathsf{T}\,\mathbf{u}}\Big)^{-1}\,\mathbf{u} \tag{2.5}$$

$$= \begin{pmatrix}\Gamma_x^{-1}(\mathbf{u})\\ \Gamma_y^{-1}(\mathbf{u})\end{pmatrix} = \begin{pmatrix}\frac{2\,u}{1+\sqrt{1-4\,\eta\,(\,u^2+v^2)}}\\ \frac{2\,v}{1+\sqrt{1-4\,\eta\,(u^2+v^2)}}\end{pmatrix}. \tag{2.6}$$

The mapping $\mathbf{\Gamma}^{-1}$ consists in shifting points towards the center and along the radial directions. The amount of shifting increases with the distance of the point to the image center (the radius). Given that the radius of $\mathbf{x}$ is $r = \sqrt{\mathbf{x}^\mathsf{T}\mathbf{x}}$, the corresponding undistorted radius is

$$r^u = (1 + \eta r^2)^{-1} r. \tag{2.7}$$

Henceforth, and in order to make the compression undergone by a particular image

more intuitive, the amount of distortion will be quantified by

$$\%\,\text{RD} = \frac{r_M^u - r_M}{r_M^u} \times 100 = -\eta r_M^2 \times 100 \tag{2.8}$$

with r_M being the distance from the center to an image corner (maximum distorted radius).

2.3 SIFT Performance in Radial Distorted Images

Mikolajczyk *et al.* [41] evaluate and compare several techniques for keypoint detection and matching under different imaging conditions including transformations in scale and rotation, affine viewpoint changes, image compression, and variation in the illumination [41, 45, 46]. The current section extends this study for the case of SIFT detection and matching in images with radial distortion. The tests are run on a set of real images that are warped using the mapping of Eq. 2.6. As explained in the introduction, the synthetic addition of geometric deformation enables fully controlled experiments with reliable ground truth, and assurance that the observations are only due to the distortion effect. The alternative would be to acquire images from the exact same viewpoint using cameras with different amounts of lens distortion. However, this is difficult to achieve in practice and small shifts in camera position, or other changes in image acquisition conditions, can potentially influence the final measurements. It is true that the interpolation in the warping can also cause undesired interferences but, as we will see latter, the final experiments with real distorted images confirm the conclusions drawn in this section.

2.3.1 Measuring Detection Performance

Consider an image $\mathtt{I}_0$ from the test set and its distorted version $\mathtt{I}_d$ with $\%RD = d$ (see Fig. 2.2). Let S_0 and S_d be respectively the sets of keypoints detected in $\mathtt{I}_0$ and $\mathtt{I}_d$. If the detection is invariant to RD, then $S_0 = S_d$ meaning that the algorithm finds the exact same points independently of the amount of deformation present in the images. Unfortunately the non-linear distortion modifies the image spectrum and SIFT does

(a) $\mathtt{I}_0(\%RD = 0\%)$ (b) $\mathtt{I}_5(\%RD = 5\%)$ (c) $\mathtt{I}_{15}(\%RD = 15\%)$ (d) $\mathtt{I}_{35}(\%RD = 35\%)$

Figure 2.2: The performance is evaluated on a data set comprising of 20 640×480 images collect on the internet with different types of visual contents.The radial distortion is artificially added by warping each image using the mapping of Eq. 2.4. The figure shows one of the images of the data set to which is added increasing amounts of deformation.

not satisfy this invariance property. The set S_d can be divided into two subsets: the set S_d^{true}, that comprises the keypoints that are simultaneously detected in the distorted and undistorted images

$$S_d^{true} \;=\; S_d \cap S_0\,, \tag{2.9}$$

and the set S_d^{false} that contains the points in $\mathtt{I}_d$ that have no correspondence in $\mathtt{I}_0$

$$S_d^{false} \;=\; S_d - S_d^{true}$$

A keypoint in the distorted image belongs to the set S_d^{true} *iff* there is a detection in $\mathtt{I}_0$ that is consistent both in space and in scale[1]. The consistency in space is verified using the mapping of Eq. 2.6. If a keypoint is detected at location $\mathbf{x}$ in image $\mathtt{I}_d$, then there must exist a keypoint in image $\mathtt{I}_0$ at location $\mathbf{u}$. In addition the scales at which the two keypoints are detected must agree. If the keypoint in the distorted image has scale σ_d, then the keypoint in the undistorted image must have scale

$$\sigma_0 \;=\; \frac{\sigma_d}{1+\eta r^2}\,, \tag{2.10}$$

[1]We follow the criteria proposed in [45, 46] according which the consistency in space and scale implies that the overlap error between keypoint regions is less than 30%. However, instead of counting all region pairs with an overlap above 70%, we only consider the pair with smallest error in order to assure one-to-one correspondence [13, 47].

where r denotes the keypoint radius in $\mathtt{I}_d$. As shown in Fig. 2.2, the addition of radial distortion diminishes the size of the image features. The evaluation takes into account this effect by performing an adaptive correction of scale using a local linear approximation of the distortion function.

In the set S_d^{false} we can distinguish between keypoints that have a match in the undistorted image $\mathtt{I}_0$ in terms of space location but not in terms of scale, and the keypoints that have no correspondence at all in S_0. The former define the subset S_d^{ws} of detections at a *wrong scale* , while the latter define the subset S_d^{new} of *newly* detected points.

$$S_d^{false} \;=\; S_d^{ws} \;\cup\; S_d^{new}$$

The subsets discussed above are used to establish different metrics for characterizing the SIFT detection. The repeatability for a certain amount of distortion $RD = d$ is computed as

$$\%_{\text{Repeatability}} = \frac{\#S_d^{true}}{\#S_0} \times 100\,, \tag{2.11}$$

with $\#$ denoting the cardinality of the set. The occurrence of new spurious detections due to the effect of radial distortion is quantified by:

$$\%_{\text{New detections}} = \frac{\#S_d^{new}}{\#S_d} \times 100\,.$$

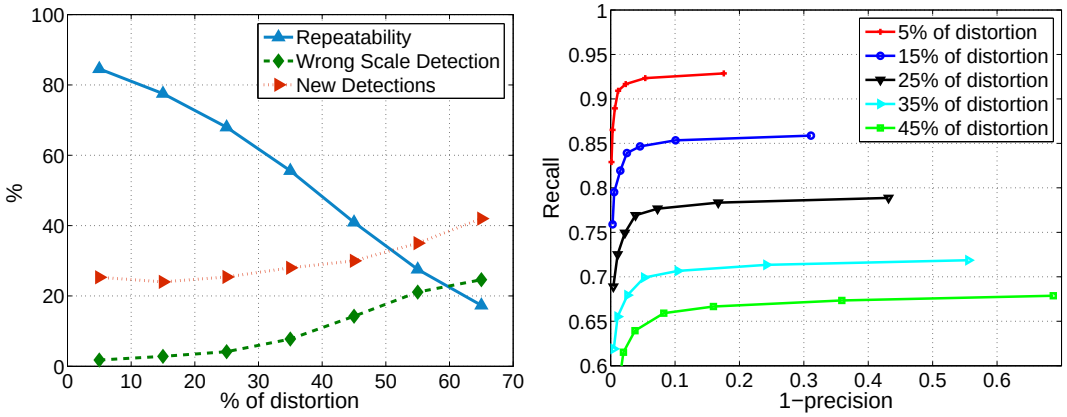
Finally, the detection at wrong scale is characterized by:

$$\%_{\text{Keypoints at wrong scale}} = \frac{\#S_d^{ws}}{\#(S_d - S_d^{new})} \times 100$$

The graphic of Fig. 2.3(a) shows the SIFT detection performance when the radial distortion increases. The measurements are obtained by averaging the results for all the images in the data set.

2.3.2 Measuring Matching Performance

Assume again an image $\mathtt{I}_0$ and one of its distorted versions $\mathtt{I}_d$. Two keypoints are considered to be a match *iff* the euclidean distance between their SIFT descriptors



(a) Detection Performance (b) 1-precision *vs* recall

Figure 2.3: SIFT detection and matching in images with radial distortion. The experimental evaluation is carried by adding an increasing amount of RD to the images of Fig. 2.2. The graphic on the left concerns detection repeatability, while the graphic on the right shows matching precision-recall curves.

is below a certain threshold λ [3]. Let M_d be the set of keypoints in $\mathtt{I}_d$ for which the matching algorithm finds a correspondence in $\mathtt{I}_0$. The elements of M_d can be divided into correct matches M_d^{true}, and incorrect matches M_d^{false}. In the best case the number of correct correspondences equals the number of correct detections. Thus, the ability of the matching algorithm in finding correct matches can be quantified using the following metric:

$$\text{recall} \;=\; \frac{\#M_d^{true}}{\#S_d^{true}}\,.$$

The *recall* must be complemented by the *precision* that measures how well the algorithm discards keypoints that have no correspondence

$$\text{precision} \;=\; \frac{\#M_d^{true}}{\#M_d}\,. \tag{2.12}$$

The *precision* and the *recall* depend on the value of the threshold λ. In general a good matching performance is achieved whenever there is a choice for λ that makes both the *precision* and the *recall* close to 1. Thus, and in a similar manner to what is done in [31, 32], the matching can be evaluated by verifying if the curve *1-precision Vs.*

recall for varying λ passes at a short distance of the operation point $(0, 1)$. Fig. 2.3(b) plots these curves for different amounts of radial distortion, with each curve being obtained by averaging the results for all the images in the data set.

2.3.3 Discussion of the Results

This section tries to interpret and understand the results observed in Fig. 2.3. From Fig.2.3(a) it follows that the repeatability of SIFT detection is severely affected by RD. There are points in the original image that are no longer detected in the images with distortion, and there are other points that, despite of being correctly located, are assigned with an incorrect scale. We observed experimentally that for increasing values of RD the keypoint detections tend move downwards in the `DoG` pyramid. This is explained by the fact that the distortion compresses the image structures and diminishes their characteristic length. Therefore, many keypoints with finer scales vanish in the presence of distortion (missing keypoints), while other keypoints with coarser scales give raise to extrema in the `DoG` pyramid that occur at lower levels than they would occur in the absence of distortion (wrong scale detections).

Fig. 2.3(a) also shows that the distortion generates a significant number of new keypoints. This is due to the fact that RD adds unstable high frequency components to the image spectrum. The SIFT detection applies fixed size Gaussians for computing each scale of the `DoG` pyramid. Since the distortion compresses the visual structures in the image periphery, the Gaussians select contributions that were not present in the original undistorted image, which gives raise to unstable keypoint detections.

Fig. 2.3(b) shows the curves of *1-precision Vs. recall* for the matching between original images and their distorted versions. The curves pass further away from the ideal operation point $(0, 1)$ as the value of added distortion increases. The RD affects the matching performance because it modifies the SIFT descriptors in two ways: First, the shift of the image pixels towards the center and along the radial direction causes a change in the image gradients. This affects the histograms that are used to build the descriptor vector (section 2.2.1). Second, the Gaussian weighting of the contributions looses its efectiveness. As the distortion increases, pixels in the periphery of the description region move closer to the keypoint, and contributions that would be negli-

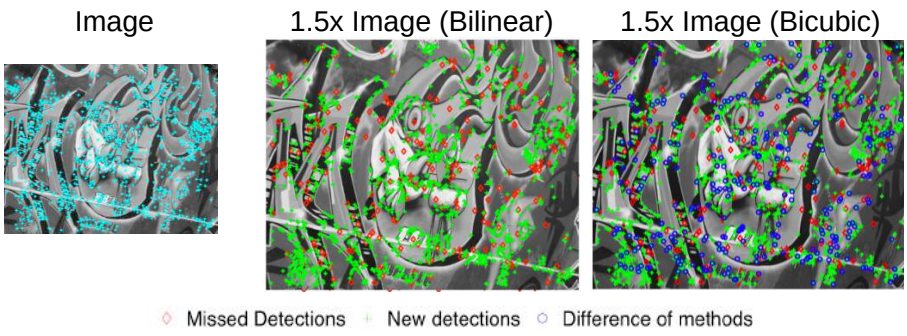


Figure 2.4: SIFT detection in re-sampled images. The size of the left-most image is increased in 50% using bi-linear and bi-cubic interpolation. SIFT keypoint detection is carried independently in each frame and the results are compared. The reasons for new detections are explained in [3]. More surprisingly is the fact that there are keypoints in the original image that are not detected in the expanded versions. It can also be observed that the detection results depend on the interpolation that is used.

gible in the absence of RD tend to become significant. In summary, the matching fails because the distortion modifies the SIFT vector, moving it away from the undistorted SIFT vector in the description space.

2.4 Keypoint Detection in Images with RD

The evaluation above shows that the repeatability of keypoint detection decreases in the presence of significant radial distortion. This section proposes strategies for overcoming the problem. We start by discussing the benefits and drawbacks of using explicit image warping for correcting the distortion. Section 2.4.2 derives a new adaptive filter that compensates for the distortion while building the image scale-space representation. Section 2.4.3 shows that the adaptive kernel can be approximated by a filter that is separable in X and Y which enables improving the computational efficiency. Finally, the new keypoint detector is evaluated and characterized in section 2.4.4.

2.4.1 Explicit Distortion Correction using Image Warping

The radial distortion causes a non-uniform compression of the image structures that affects SIFT performance. Keypoints at finer scales vanish, others are detected at

lower scales than they would be in the absence of distortion, and there are new unstable detections in the image periphery due to spurious high-frequency components introduced by RD (see Fig. 2.3 and 2.8). A possible strategy for avoiding this compressive effect is to explicitly correct the distortion by image warping and detect the keypoints in the DoG pyramid of the rectified image [33]. This approach, henceforth dubbed *rectSIFT*, is evaluated in Fig. 2.8.

In a first analysis we would expect a detection repeatability close to 100%. However, and despite of the significant improvements with respect to standard SIFT, the detection results are far from this score. The distortion correction by image re-sampling implicitly requires reconstructing the signal from the initial discrete image. The problem is that, not only there are high frequency components that can not be recovered (e.g low resolution, aliasing), but also the reconstruction filters are imperfect. The bi-linear and bi-cubic interpolations are respectively first and second order approximations of the ideal reconstruction kernel that is the infinite *sinc* function [34]. Such approximations introduce spurious frequency components and other signal artifacts that affect the keypoint detection. The skeptical reader can easily verify this by observing Fig. 2.4. The left-most image is linearly re-scaled by a factor of 1.5, and SIFT keypoint detection is ran both in the original and expanded images. Remark that, since the signal resolution is increased, there are neither aliasing effects nor losses of high-frequency components. We would expect for the scale invariant detector to find in the expanded images all the keypoints detected in the original frame. This clearly does not happen. Moreover the detection results depend on the type of interpolation that is used to perform the re-scaling.

In summary, as first argued by Daniilidis et al. [29], explicit distortion correction by image warping should be avoided because interpolation effects introduce undesired results in filtering. This largely explains the evaluation results shown in Fig. 2.8. It is curious to observe that for distortions below 15% the standard SIFT detection outperforms rectSIFT. It means that for small amounts of RD the pernicious effects of image re-sampling surpass the benefits of correcting the radial distortion.

2.4.2 Adaptive Gaussian Filtering

We propose a model-based approach for image blurring that compensates for the spectral modifications caused by radial distortion. While in rectSIFT the DoG pyramid is computed after warping the image, in this section the scale-space representation is generated directly from the frame with distortion using adaptive Gaussian filtering. The outcome is a DoG pyramid equivalent to the one that would be obtained by following the steps:

1. Correct the radial distortion of the image $\mathtt{I}$
2. Blur $\mathtt{I}^u$ through successive convolutions with a Gaussian function.
3. Apply radial distortion to the blurred images $\mathtt{L}^u$
4. Subtract the distorted blurred images L for obtaining the final DoG pyramid

As we will see latter, the detection repeatability improves dramatically by avoiding the image re-sampling required by the warping operation. The adaptive Gaussian function is derived below.

Consider the convolution of the undistorted image $\mathtt{I}^u$ with a Gaussian kernel with standard deviation σ. By writing the convolution operation of Eq. 2.1 explicitly, it comes that the blurred image is

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

If I is the original image with distortion, then it follows from section 2.2.2 that $\mathtt{I}^u(\mathbf{u}) = \mathtt{I}(\mathbf{x})$ with $\mathbf{x} = \mathbf{\Gamma}^{-1}(\mathbf{u})$ (equation 2.6). Replacing $\mathtt{I}^u$ by I and switching the variables (u, v) by (x, y) using the mapping relation 2.6, we obtain the result of Eq. 2.13

$$\mathtt{L}^u_\sigma(s,t) \;=\; \sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} \mathtt{I}(x,y)\, \mathsf{G}_\sigma\Big(s-\Gamma_u\,(x,y),\, t-\Gamma_v\,(x,y)\Big)\,. \tag{2.13}$$

This equation computes the undistorted blurred image $\mathtt{L}^u$ directly from the original distorted frame I. However, it is no longer is a strict convolution because the filter

function varies with the image location that is being filtered. Henceforth, we will refer to this operation as being an *adaptive convolution* that is denoted by $\star$ whenever convenient.

Let's now apply radial distortion to the blurred image $\mathtt{L}^u$ in order to obtain $\mathtt{L}$. This can be achieved in an implicit manner using again the mapping relations of section 2.2.2. Equation

$$\mathtt{L}_\sigma(h,k) \;=\; \sum_{x=-\infty}^{\infty}\sum_{y=-\infty}^{\infty} \mathtt{I}(x,y)\,\mathsf{G}_\sigma\Big(\Gamma_u\,(h,k)-\Gamma_u\,(x,y),\,\Gamma_v\,(h,k)-\Gamma_v\,(x,y)\Big) \quad (2.14)$$

is derived by a new switching of variables, that substitutes the undistorted image coordinates (s,t) by their distorted counterpart (h,k). After replacing $\mathbf{\Gamma}^{-1}$ and performing some algebraic simplifications, we obtain the adaptive filtering of Eq. 2.15

$$\mathtt{L}_\sigma(h,k) \;=\; \sum_{x=-\infty}^{\infty}\sum_{y=-\infty}^{\infty} \mathtt{I}(x,y)\,\mathsf{G}_\sigma\Big(\frac{h-x+\eta r^2(h\delta^2-x)}{1+\eta r^2(1+\delta^2+\eta r^2\delta^2)},\frac{k-y+\eta r^2(k\delta^2-y)}{1+\eta r^2(1+\delta^2+\eta r^2\delta^2)}\Big) \quad (2.15)$$

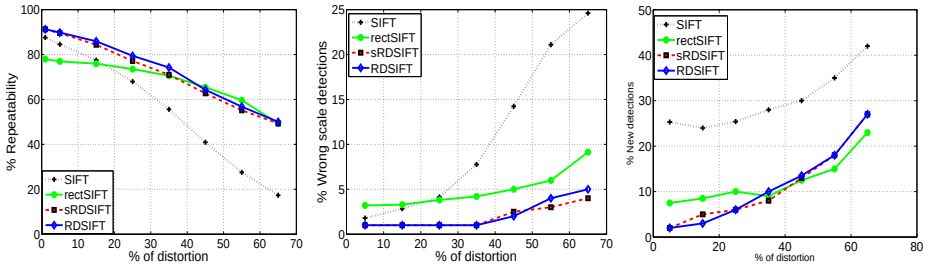
with r being the distance between the center and the image location where the filter is applied

$$r \;=\; \sqrt{h^2+k^2}\,, \quad (2.16)$$

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

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

The keypoints are detected by looking for extrema in the $\mathtt{DoG}$ pyramid that is computed by subtracting the images $\mathtt{L}$ of Eq. 2.15 for increasing values of σ (see Eq. 2.2. The application of this adaptive filtering is called the RD-SIFT algorithm. Figure 2.8 shows that the RD-SIFT outperforms the standard SIFT in every evaluation parameter and level of distortion. More importantly, RD-SIFT is unarguably better than rectSIFT for amounts of distortion up to 45%. Beyond this point the compressive effect is so strong that many image structures disappear and can no longer be filtered out. Since rectSIFT tries to restore the original signal, it tends to provide slightly better repeatability under very extreme RD. However, not only this relative superiority is



(a) Detection Repeatability (b) Detection at Wrong Scales (c) Detection of New Keypoints

Figure 2.5: The graphics compare the performance of different strategies in detecting keypoints in images with radial distortion. The detection methods are: standard SIFT applied to original distorted images (SIFT); standard SIFT applied to frames where the distortion has been corrected using explicit image warping (rectSIFT); search for extrema in a DoG pyramid obtained using the adaptive Gaussian kernel derived in section 2.4.2 (RD-SIFT); and search for extrema in a DoG pyramid obtained using the separable approximation of the adaptive kernel derived in section 2.4.3 (sRD-SIFT). The evaluation is carried using the images and methodology of section 2.3.

almost negligible, but also such high amounts of distortion are unlike to arise in real camera systems.

2.4.3 Improving Computational Efficiency

The adaptive convolution of Eq. 2.15 is computationally intensive both in terms of processing and memory requirements [29]. We now discuss an approximation of the filter function that enables conciliating good detection repeatability with computational efficiency. Let's analyze how the filter of Eq. 2.15 adapts to the RD present in the image. Consider that the image point with coordinates (h, k) is near the center. In this case the term ηr^2 is very close to zero and the filtering operation converges to the standard convolution by a Gaussian kernel. This makes sense because, since the effect of RD is usually unnoticeable in the center, there is no need for the filter to make any type of compensation. Consider now that the point (h, k) is in the image periphery. Since the filtering kernel dismisses pixel contributions far away from the convolution center, it is reasonable to assume that (x, y) is close to (h, k) and that the ratio δ is

approximately unitary. Making $\delta = 1$ in Eq. 2.15 yields

$$\widehat{\mathtt{L}}_\sigma(h,k) = \sum_{x=-\infty}^{\infty}\sum_{y=-\infty}^{\infty} \mathtt{I}(x,y)\,\mathsf{G}_\sigma\Big(\frac{h-x}{1+\eta r^2},\,\frac{k-y}{1+\eta r^2}\Big)\,,$$

with $\widehat{\mathtt{L}}$ being an approximation of $\mathtt{L}$. The expression can be re-written using the adaptive convolution operator:

$$\widehat{\mathtt{L}} = \mathtt{I} \star \widehat{\mathsf{G}} \tag{2.17}$$

where $\widehat{\mathsf{G}}$ is given by

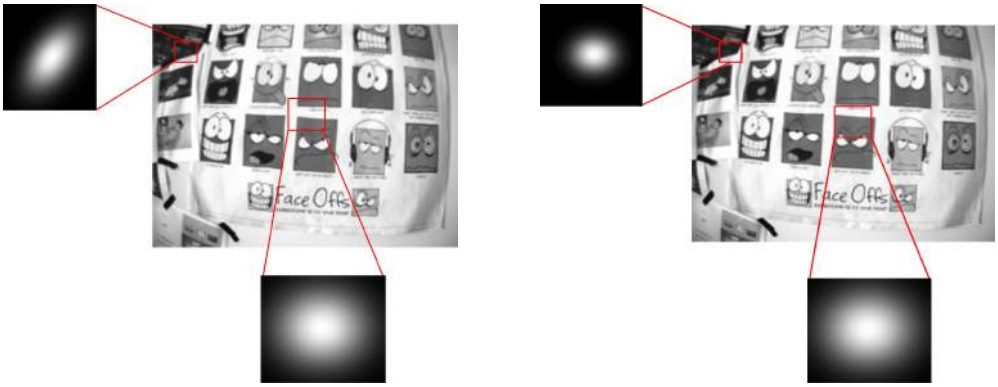
$$\widehat{\mathsf{G}} = \mathsf{G}_{\sigma'}(x,y), \tag{2.18}$$

with $\sigma' = (1 + \eta\, r^2)\sigma$. From the Eq. 2.18 it is easy to understand that $\mathtt{I}$ is filtered by a Gaussian kernel with a standard deviation that varies with the image radius r. As we move far from the center, the filter adapts to the distortion by increasingly emphasizing the pixel contributions closer to the convolution point. While the filtering of Eq. 2.15 uses a different kernel at every image pixel location, the approximation of Eq. 2.17 employs the same filter function for image locations equidistant to the center. This decrease in the number of kernels is advantageous for implementations using a look-up table of pre-computed filter masks for speeding up the convolution process. Refer to Fig. 2.6 for a comparison between the accurate and this simplified filter.

It is well know that the regular 2D Gaussian function G can be generated by cascading two 1D Gaussian kernels [34]. This decoupling property is used in standard scale-space implementations for dramatically decreasing the computational complexity of image blurring. The filtering is typically achieved by successively convolving the image with a 1D Gaussian kernel with horizontal and vertical orientations. Unfortunately, neither the exact filter of Eq. 2.15, nor $\widehat{\mathsf{G}}$, verify the decoupling property. Despite of this let's consider the adaptive kernel $\mathring{\mathsf{G}}$ given by

$$\mathring{\mathsf{G}} = \mathbf{g}^h_{\sigma'}(x,y) \star \mathbf{g}^v_{\sigma'}(x,y)\,, \tag{2.19}$$

with $\mathbf{g}^h_{\sigma'}$ and $\mathbf{g}^v_{\sigma'}$ being horizontal and vertical 1D Gaussian functions with standard deviations varying with the radius of the convolution center.



(a) RD-SIFT Gaussian filter (b) sRD-SIFT Gaussian filter

Figure 2.6: Differences between accurate and simplified filters. Tha RD-SIFT Gaussian filters change their shape according to the pixel radius and orientation according to the pixel position in the image. The filter is strongly non-isotropic and non-separable. The sRD-SIFT filters only vary their shape according to the pixel radius. Since they are spatial invariant with respect to the image radius, this allows to perform separable convolution in X and Y dimensions by cascading two 1D Gaussian filters.

Figure 2.7 studies how well $\mathring{\mathsf{G}}$ and $\widehat{\mathsf{G}}$ approximate the filtering of Eq. 2.15. The three adaptive kernels are generated for every possible image location, and the similarity between them is evaluated using normalized cross correlation (NCC). The first column shows the NCC scores between $\mathring{\mathsf{G}}$ and the exact filter. It can be observed that the approximation becomes worse when the RD increases and the convolution center moves towards the image periphery. However, for most image pixel locations the two kernels are quite similar. The second column depicts the NCC for $\widehat{\mathsf{G}}$ and the behavior is in general the same. The third column is obtained by subtracting the result of the second column to the first. It is interesting to observe that, since the difference in NCC is always positive, the adaptive kernel $\mathring{\mathsf{G}}$ is slightly better than $\widehat{\mathsf{G}}$ in approximating the exact filter function.

Summarizing, the filtering of Eq. 2.15 can be approximated by an adaptive convolution using either $\mathring{\mathsf{G}}$ or $\widehat{\mathsf{G}}$. The former approximation is better than the latter and, more importantly, the filtering can be implemented by cascading two 1D adaptive Gaussian filters with horizontal and vertical orientations. Fig. 2.8 evaluates the performance of the *sRD-SIFT* algorithm that implements the image blurring in a decoupled manner.

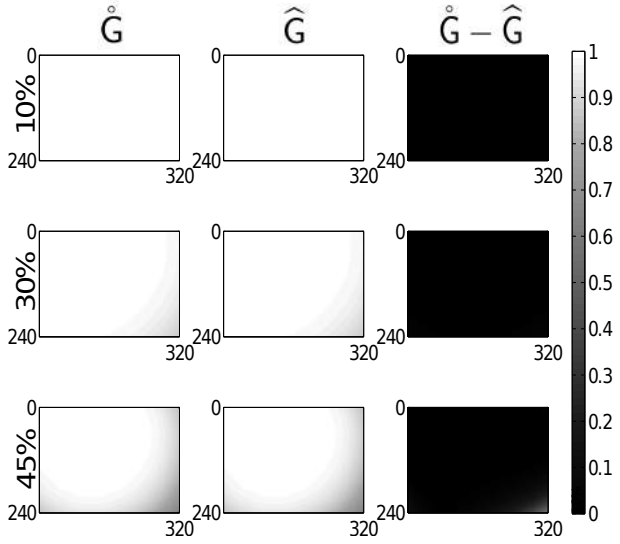
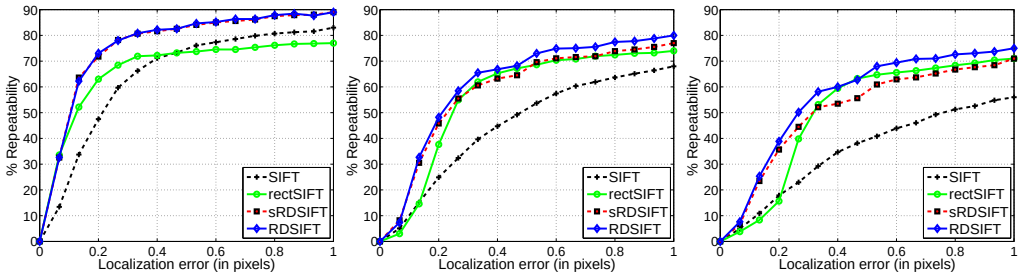


Figure 2.7: The figure shows how well the kernels $\mathring{\mathsf{G}}$ and $\widehat{\mathsf{G}}$ approximate the filtering of Eq. 2.15. The rectangles represent the lower right quadrant of a 640×480 image with radial distortion (the RD value is different for each row). The two approximation kernels and the exact adaptive filter are generated for every possible image pixel location. In the first column a color scale is used to show the normalized coss-correlation (NCC) score between the exact filter mask and $\mathring{\mathsf{G}}$. The second columns does the same for $\widehat{\mathsf{G}}$. The third column depicts the difference of the two first columns that is multiplied by 1000 for visualization purposes.

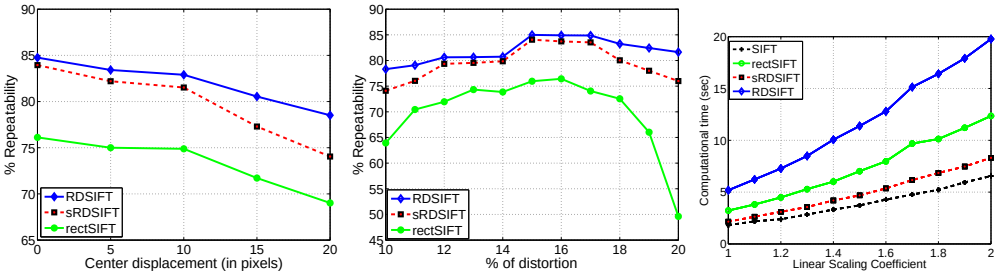
The images $\mathring{\mathsf{L}}$, that give raise to the DoG pyramid, are obtained by the convolution of the original image I with the 1D filters $\mathbf{g}^h_{\sigma'}$ and $\mathbf{g}^v_{\sigma'}$. The adaptive filters are pre-computed and stored in a look-up table enabling an implementation with an overall computation performance very close to standard SIFT. As expected, the approximated filtering used in sRD-SIFT causes a slight decrease in detection performance when compared to RD-SIFT. However, the deterioration of detection is in general small, and largely compensated by the improvements in computational efficiency (see Fig. 2.8(d)).

2.4.4 Additional Evaluations

In this section we run some additional tests to better evaluate and compare the detection performances of SIFT, rectSIFT, RD-SIFT, and sRD-SIFT. All the experiments are carried using the data set introduced in section 2.3. The RD is added artificially



(a) Sub-pixel accuracy in keypoint detection. The graphics show the repeatability when the tolerance in position error increases. From left to right the RD is 5%, 25%, and 35%.



(b) Disturbance in the Center (c) Disturbance in % of RD (d) Time Profiling

Figure 2.8: The figures concern the spatial accuracy of keypoint detection (a), the robustness to errors in the calibration parameters (b)-(c), and the computation time for building the DoG pyramid (d). The results of (a)-(c) were obtained using the images of Fig. 2.2. The tests on sensitivity to calibration parameters were carried assuming $RD = 15\%$. The computational time of (d) was evaluated for images (640×480) with increasing size and a constant distortion of 25%.

and the SIFT detections in the original undistorted image are used as ground-truth.

2.4.4.1 Sub-pixel Accuracy in Keypoint Detection

Recent works assess sub-pixel detection accuracy by evaluating the repeatability using different position error thresholds for deciding about the keypoint correctness [13,47]. We follow a similar methodology and show in Fig. 2.8 the repeatability curves obtained by varying the tolerance in the location error from 0 to 1 pixel. It can be observed that the increase in RD affects both the repeatability and the accuracy of the keypoint localization. The original SIFT algorithm is the one where the break in accuracy is more pronounced, while RD-SIFT is the method more resilient to the deterioration. It is interesting to observe that for strong distortion (RD=35%) there is a

scarce number of rectSIFT detections with accuracy below 0.2 pixels. sRD-SIFT has a behavior similar to RD-SIFT for low amounts of distortion, but the break in accuracy is more noticeable when RD increases. This can be easily understood by taking into account that the approximation in the filtering becomes coarser. Comparing sRD-SIFT with rectSIFT we might conclude that in general the former is significantly more accurate than the latter.

2.4.4.2 Robustness to calibration errors

The algorithms RD-SIFT, sRD-SIFT, and rectSIFT, require prior knowledge about the center and amount of distortion. In this experiment we evaluate the robustness of the detection to noise in the calibration parameters. Fig. 2.8(b) shows the repeatability behavior when the position error in the distortion center ranges from 0 to 20 pixels (the shift direction is random). As expected, all the methods are affected by inaccurate center calibration, but the break in performance is smooth and proportional to the disturbance. The behavior of the three algorithms is very similar, with RD-SIFT being slightly more robust than the competitors. Fig. 2.8(c) shows the repeatability when the error is in the quantification of the RD. Both RD-SIFT and sRD-SIFT present a reasonable robustness to the disturbance (the former more than the latter). rectSIFT seems to be more sensitive, specially when the RD is over-estimated. We believe that this is due to a poorer image signal reconstruction because of the wider interpolation intervals. From the study we can say that the proposed algorithms lead to significant improvements in detection repeatability, even when the RD calibration is performed in a coarse manner.

2.4.4.3 Run time

This experiment compares the execution time of the different detectors with respect to increasing image resolution. Fig. 2.8(d) shows the average run time on the images of the data set after proper scaling and addition of RD=25%. The measured *detection time* is the sum of the time intervals spent in pre-processing, generating the scale-space representation, and looking for local extrema. In RD-SIFT and sRD-SIFT the pre-processing consists in computing the adaptive filter masks and storing them into

memory, while in rectSIFT it refers to correcting RD through image re-sampling [2]. From Fig. 2.8(d) it follows that sRD-SIFT has a computational efficiency close to standard SIFT. We verified experimentally that the overhead introduced by the adaptive filtering is usually negligible, and that the time difference is caused by the pre-processing step. The graphic also shows that rectSIFT is substantially less efficient, presenting an execution time that grows exponentially with the image resolution. The overhead comes from the interpolation in the pre-processing stage and from the larger size of the undistorted frame. Since the RD correction expands the image, the filtering and looking for extrema become computationally more expensive.

2.5 Keypoint Description in Images with RD

The SIFT description is not invariant to RD because the non-linear deformation changes pixel positions and image gradients in the neighborhood of the keypoint. As a consequence, the SIFT vector is displaced in the description space with respect to its position in the absence of RD. Since the RD deformation is non-uniform across the image, the descriptor displacement depends on the location where the keypoint is detected. This variability precludes using any kind of nearest-neighbors strategy for successfully matching keypoints across different views (see Fig. 2.9(a)). This section shows how to keep the descriptor vector stationary in order to achieve RD invariance.

2.5.1 Implicit Gradient Correction

The most straightforward approach to achieve RD invariance on the description step would be through explicitly rectification, by warping the image and computing the gradients in the undistorted signal. Since we aim at working at with the original pixel values, we perform an implicit correction by measuring the gradients in the original image and correct them using a derivative chain rule. The implicit approach avoids the propagation of interpolation artifacts inherent to the image re-sampling, and is

[2] Remark that, for the case of detection in an image sequence, the explicit RD correction must be repeated for each frame, while the adaptive masks are computed only once.

computationally more efficient because the gradient correction is only performed in the description regions around the keypoints.

Let $\mathtt{I}$ be the original image and $\mathtt{I}^u$ be its undistorted counterpart. From section 2.2.2 it follows that

$$\mathtt{I}^u(\mathbf{u}) \;=\; \mathtt{I}(\mathbf{\Gamma}^{-1}(\mathbf{u}))\,.$$

Applying the derivative chain rule it yields

$$\nabla \mathtt{I}^u \;=\; \mathsf{J}_{\mathbf{\Gamma}^{-1}}\,.\,\nabla \mathtt{I} \tag{2.20}$$

with $\nabla \mathtt{I}^u$ and $\nabla \mathtt{I}$ being respectively the gradient vectors in $\mathtt{I}^u$ and $\mathtt{I}$, and $\mathsf{J}_{\mathbf{\Gamma}^{-1}}$ being the 2×2 jacobian matrix of the mapping function $\mathbf{\Gamma}^{-1}$ given in Eq. 2.6. The Jacobian matrix can be written in terms of distorted image coordinate $\mathbf{x} \;=\; (x, y)^{\mathsf{T}}$ by replacing $\mathbf{u}$ using the mapping of Eq. 2.4:

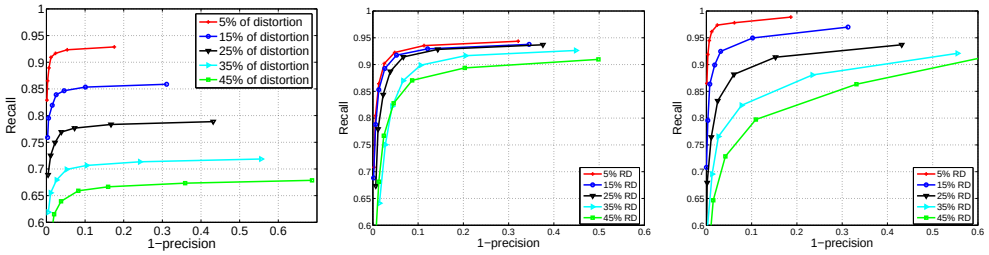
$$\mathsf{J}_{\mathbf{\Gamma}^{-1}} \;=\; \frac{1+\eta r^2}{1-\eta r^2} \begin{pmatrix} 1-\eta(r^2-8x^2) & 8\eta xy \\ 8\eta xy & 1-\eta(r^2-8y^2) \end{pmatrix}$$

with r denoting the radius of $\mathbf{x}$.

In summary, we propose to measure the gradients directly in the original distorted image $\mathtt{I}$, evaluate the Jacobian matrix $\mathsf{J}_{\mathbf{\Gamma}^{-1}}$ at every relevant pixel location, and correct the gradient vectors $\nabla \mathtt{I}$ using Eq. 2.20. The keypoint descriptor is generated from the undistorted gradients $\nabla \mathtt{I}^u$ following the procedure described in 2.2.1. The only modification is the replacement of the weighting Gaussian function $\mathsf{G}(x, y; \sigma)$ by the function $\widehat{\mathsf{G}} \;=\; \mathsf{G}_{\sigma'}(x, y)$ that accounts for the changes in pixel contributions due to RD.

2.5.2 Evaluation in Keypoint Matching

Fig. 2.9 shows the precision-recall for keypoint matching using SIFT descriptors generated before and after compensating for the image distortion. The comparison with standard SIFT description shows a dramatic improvement in the retrieval performance. Thus, the first conclusion is that the correction of image gradients enables achieving



(a) RD images (SIFT) (b) Explicit Correction (rect-SIFT) (c) Implicit Gradient Correction

Figure 2.9: Matching performance by generating the SIFT descriptors before (2.9(a)) and after compensating for the distortion (2.9(b) and 2.9(c)). The graphics show the curves of *1-precison Vs. recall* for increasing amounts of RD. In Fig. 2.9(b) the distortion is explicit corrected via image warping, while in Fig. 2.9(c) we use the implicit gradient correction approach described in 2.5.1.

RD invariance during description which boosts the keypoint matching performance. By comparing implicit gradient correction against explicit image warping, it comes that the former is superior to the latter for amounts of distortion up to 25%. This is explained by the fact that the interpolation employed in the re-sampling process introduces spurious frequency components that propagate for the first order derivatives that are used in the descriptor vector. For very strong distortions the explicit image rectification outperforms the implicit gradient correction. As discussed previously, beyond a certain amount of RD the compressive effect becomes so strong that local variations that would be observed in the undistorted image are no longer detectable in the distorted signal. In other words, it is impossible to recover the gradient vector $\nabla \mathtt{I}^u$ using Eq. 2.20 because the corresponding vector $\nabla \mathtt{I}$ cannot measured. In this case the interpolation used in the re-sampling is advantageous because it enables inferring missing information.

2.6 Experimental Validation

This section aims to confirm the results so far by running experiments in images acquired by real cameras with lens distortion that undergo changes in scale, rotation, and viewpoint. The sRD-SIFT keypoint detection and matching is compared against

the original SIFT algorithm, the SIFT run after performing explicit RD correction via image warping (rectSIFT), and the pSIFT framework [32]. As discussed in section 2.1, the pSIFT detection approximates the spherical diffusion using a stereographic projection and computes the descriptor by considering a support region on the sphere, which is re-sampled to a canonical patch of size 41×41.

Figure 2.10: Calibration grids and 3 (out of 13) images for each data set used for the experiments of section 2.6.1. The frames were acquired using a lens with low distortion (RD$\approx$ 10%), a 4mm minilens commonly used for robotics' applications (RD$\approx$ 20%), and a fish-eye lens with a wide FOV (RD$\approx$ 40%). The image resolution is 640×480 for all cases.

2.6.1 Planar Textured Surfaces

This experiment uses three images sequences of planar scenes acquired using lens that introduce different amounts of distortion (see Fig. 2.10). The results of each sequence are averaged over 78 image pairs obtained from 13 frames. For the case of rectSIFT and sRD-SIFT, the distortion center is assumed to be coincident with the image center, and the distortion parameter η is roughly estimated by straightening up lines in the image periphery [44]. For the case of pSIFT, the camera intrinsics are fully calibrated from images of a checkerboard pattern using the method proposed in [48]. Since the scenes are planar, the frames are related by an homography that can be used to verify

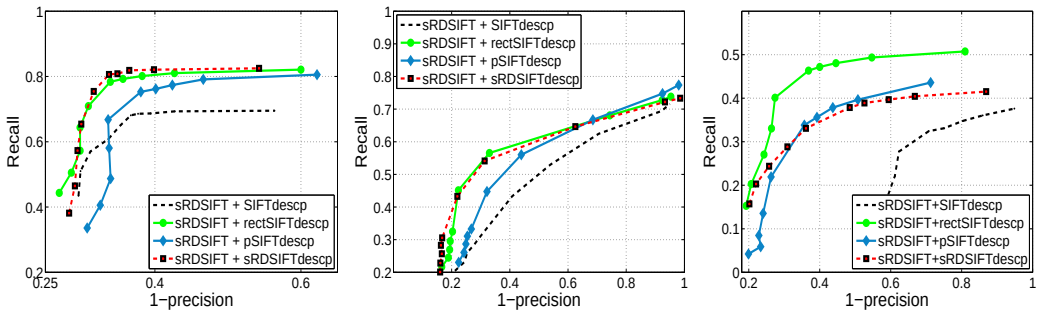
Table 2.1: The table compares the performance of the four algorithms in planar scenes. The left-most group of columns concern the computational overhead, the middle group refers to detection and matching when the threshold value for keypoint selection in the DoG pyramid is the same for all methods. $\#S_d$, $\#S_d^{true}$, and $\#M_d^{true}$ are respectively the average number of detections, of common detections in the image pair (matching potential), and of correctly established correspondences. We also show the detection repeatability and the matching precision as defined in section 2.3. The matches and respective precision values were computed using the ratio best and second neighbors descriptor thresholded at 0.8 [3]

		Time (sec)		**Detect. & Match. (Constant Threshold)**				
		Detect.	**Total**	$\#\mathbf{S}_d$	$\#\mathbf{S}_d^{true}$	$\%\mathbf{Rep.}$	$\#\mathbf{M}_d^{true}$	$\%\mathbf{Prec.}$
10%	**SIFT**	1.57	1.97	1052	596	**57**	405	**62**
	rectSIFT	3.31	3.73	1057	644	**61**	431	**70**
	pSIFT	2.05	2.78	1224	756	**61**	524	**67**
	sRD-SIFT	1.61	2.32	1080	777	**72**	528	**71**
20%	**SIFT**	1.95	2.79	1332	871	**65**	458	**47**
	rectSIFT	4.85	5.66	1375	1022	**74**	539	**68**
	pSIFT	2.21	3.37	1558	1168	**75**	654	**57**
	sRD-SIFT	1.99	3.02	1412	1110	**78**	641	**65**
40%	**SIFT**	1.87	2.35	900	295	**27**	78	**30**
	rectSIFT	18.22	20.88	752	419	**56**	165	**67**
	pSIFT	2.33	4.28	1557	795	**51**	286	**63**
	sRD-SIFT	2.01	3.98	1663	809	**49**	295	**65**

the correctness of the matches and the repeatability of detection [45, 46]. We apply a robust estimation algorithm that uses hundreds of correspondences for computing these ground truth homographies [45, 49].

Table 2.1 compares the performance of the four studied algorithms. The two left-most columns concern the computational overhead, and show the time for detection[3] and the total runtime. It can be observed that the overhead of pSIFT and sRD-SIFT with respect to the original SIFT is very small, with the former being slightly slower

[3]The detection time does not include the offline computation of the filter masks used by pSIFT and sRD-SIFT. For pSIFT the Matlab implementation supplied by the authors took around 5 minutes to compute the octave filters for each sequence. For sRD-SIFT the Matlab and C implementations took respectively 1.25 and 0.35 seconds to accomplish the task.



(a) Recall for RD= 10% (b) Recall for RD= 20% (c) Recall for RD= 40%

Figure 2.11: Keypoint matching evaluation in planar scenes with ground truth. Figures 2.11(a) to 2.11(c) depict the *1-precision Vs. Recall* curves that characterize the retrieval performance of the four descriptors being tested (in this case the keypoints were detected using sRD-SIFT)

than the latter because of the rendering of the stereographic image. In rectSIFT the exponential growth of computation time with RD is justified by the increasing size of the corrected warped frames.

The middle columns show the average results for detection and matching when the threshold for selecting keypoints in the DoG pyramid is 1.25×10^{-2}. The relative performance of SIFT, rectSIFT, and sRD-SIFT in terms of repeatability and matching precision is in accordance with the synthetic experiments of Fig. 2.8 and 2.9. For $RD = 40\%$ rectSIFT presents the highest repeatability score, but sRD-SIFT achieves substantially more detections thanks to the adaptive filtering that avoids an excessive blurring in the image periphery. Comparing sRD-SIFT with pSIFT, the former tends to achieve better repeatability and precision scores, but in overall terms the two methods behave quite similarly. Since the test images undergo significant changes in viewpoint (Fig. 2.10), the pSIFT invariance to camera rotation is an advantage that seems to compensate the drawbacks of the re-sampling used for rendering the stereographic image.

Figure 2.11 aims at comparing the four descriptors being tested. The precision-recall of each method is measured over the same set of keypoints detected using sRD-SIFT. The results are consistent with the observations made in Fig. 2.9, with the implicit gradient correction of section 2.5.1 being the top-performer for RD amounts

up to the 20% case. For very high distortion the explicit correction by interpolation provides the best keypoint description, and can be used as an alternative for further improving the matching results of our framework. Surprisingly the pSIFT descriptor presents a break in terms of descriptor distinctiveness for all levels of distortion. This fact is due to the additional re-sampling step for mapping the sphere support regions into a canonical patch of 41×41 pixels [32]. The pernicious effects of the operation might be negligible for coarse scale features, but for fine structures the interpolation intervals are often too large and induce gross errors in the rendered patch.

In summary, sRD-SIFT and pSIFT always provide the largest number of keypoints that can be correctly associated between the two frames. However, the superiority both in terms of effective number of correct matches and computation time make sRD-SIFT a better option than rectSIFT and pSIFT for low/moderate amounts of distortion. Two major advantages of sRD-SIFT over pSIFT are the fact that intrinsic camera calibration is not required (a rough approximation is enough) and also that image signal re-sampling is completely avoid by formulating all the processing steps in the image plane [29].

2.6.2 Structure-From-Motion in Medical Endoscopy

Accurate point correspondence across frames is of key importance in multiple-view geometry application, such as structure-from-motion [37, 49]. In this experiment we evaluate the different keypoint detection and matching methods in medical endoscopy structure-from-motion. The rectSIFT method is not include in the evaluation since interpolation introduces pernicious effects in terms of keypoint precision as described in [50, 51]. For computing the camera motion we use 5-point algorithm [52] in a RANdom SAmple Consensus (RANSAC) procedure that estimates the epipolar geometry in a robust manner [37]. The RANSAC is an iterative scheme that computes the essential matrix from 5 randomly chosen correspondences, and counts the number of point matches that agree with the achieved estimation. A point match is considered to be an inlier *iff* the Sampson distance to the epipolar lines is below a certain threshold value [37].

Due the nature of the rigid SfM experiments, the datasets were collected by imag-

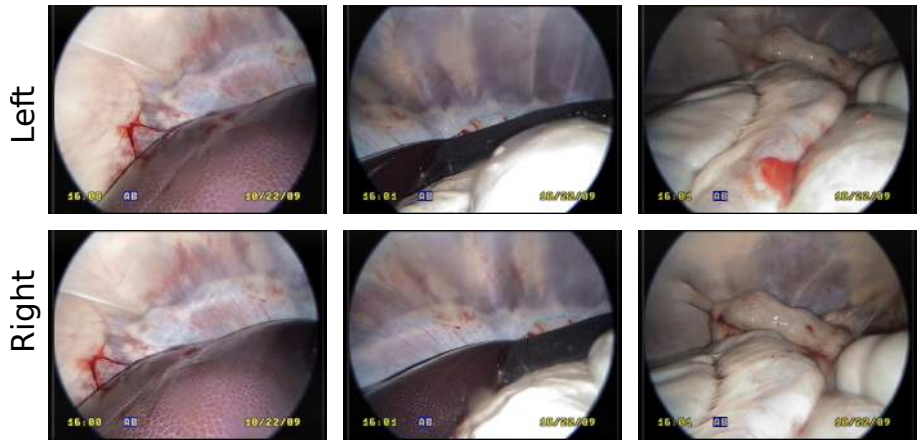


Figure 2.12: Sample images used in the medical SfM experiments.

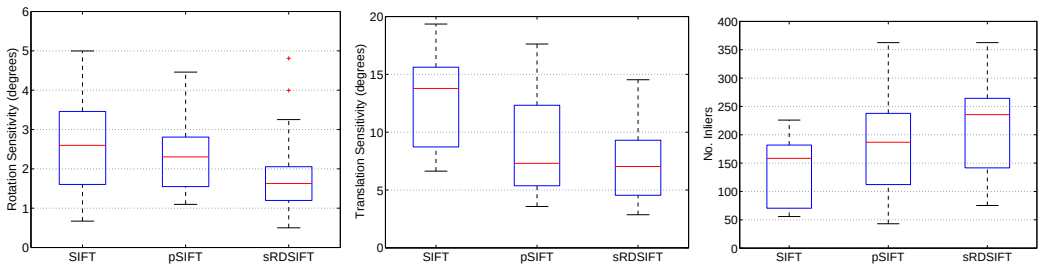
ing *ex vivo* tissues from a porcine, which enable to minimize the non-rigid physiological motions. The datasets used in this section were made available by [53, 54]. This experiment consists in computing the camera motion between 30 pairs of images with depth variation, and evaluating the repeatability of the camera motion estimates. Given $N = 50$ trials of the RANSAC plus 5-point algorithm, we compute a *mean* rotation matrix $\bar{\mathsf{R}}$ [55] and a *mean* translation vector $\bar{\mathsf{t}}$ [56], with $\bar{\mathsf{t}}$ being a unitary vector. For each image pair, the sensitivity in translation is then computed as follows:

$$\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}[\arccos(\bar{\mathbf{t}}^{\mathsf{T}}\mathbf{t}_n)]^2}. \tag{2.21}$$

Like in [56], a difference rotation matrix $\Delta\mathsf{R} = \bar{\mathsf{R}}^{\mathsf{T}}\mathsf{R}_n$ is used to compute the angular difference between the $\bar{\mathsf{R}}$ and R_n. For each image pair, the sensitivity in rotation estimates is measured by the standard deviation of the angular differences for the N RANSAC trials.

Experimental results

Figure 2.13 depicts the results using 30 images pairs. For the sake of visualization we combine the results of the 30 pairs using a boxplot, that can be used to access the variability of the estimations. It can be seen in Fig. 2.13(c) that the sRD-SIFT algorithm enables to establish more matches than SIFT and pSIFT. More important than



(a) Rotation Error (b) Translation Error (c) No Inlier Matches

Figure 2.13: Structure-from-motion in medical endoscopy. The graphics show the rotation 2.13(a) and translation 2.13(b) sensitivity analysis. The last graphic show the number of correct matches provided by each method. It can be seen that sRD-SIFT algorithm provides better camera motion estimates than the two other approaches.

the number of correct correspondences across views is their localization accuracy in terms of sub-pixel precision for recovering the camera motion. The sRD-SIFT algorithm provides the more consistent estimations for rotation and translation (see Fig. 2.13(a) and 2.13(b), respectively). The pSIFT algorithm improves upon SIFT in terms of number of matches obtained. However, the camera motion estimatives are not as consistent as the ones observed with the sRD-SIFT. We believe this is due to the extra-interpolation step required to map the image to the stereographic plane to carry feature detection. This process introduces signal artifacts that affect the keypoint precision, which propagates to the camera motion estimation. In summary we can conclude that sRD-SIFT and the pSIFT give raise to similar number of correspondences, but the sRD-SIFT enables more accurate camera motion estimations. This advantage come from the fact that sRD-SIFT completely avoids image interpolation.

2.7 Closure

This chapter proposes modifications to the broadly used SIFT framework that make it resilient to image radial distortion, while preserving the original invariance to scale, rotation, and moderate viewpoint change. The only assumptions are that the camera follows the division model [43], and that the amount of distortion is coarsely known. We ran several experiments, both in synthetic and real frames, that prove the superi-

ority of sRD-SIFT whenever there is significant image distortion. Our method often duplicates the number of correct point correspondences, while keeping a high localization accuracy. All this is achieved at the expense of a small computational overhead when compared with the standard SIFT implementation. sRD-SIFT can be advantageous in several robot vision tasks, ranging from SfM to visual recognition, as well as in medical applications that rely in endoscopic imagery.

The main virtue of proposed approach is that it avoids image re-sampling. The interpolation used in previous works that require image warping operations [32, 33] severely affects the keypoint detection performance. With sRD-SIFT we show that the radial distortion can be locally compensated using an adaptive kernel, and that this adaptive filtering can be implemented in a computationally affordable manner.

Chapter 3

Image-based Indoor Localization

One valuable competence of any robotic navigation system is the ability to localize itself with respect to the environment. This chapter investigates the problem of image-based localization in indoor environments when dealing with hybrid imaging systems, i.e. when the query image and the visual map database have been acquired with different imaging systems. The localization is achieved by querying a database of omnidirectional images that constitutes a detailed visual map of the building where the robot operates. Omnidirectional cameras have the advantage, when compared to standard perspectives, of capturing in a single frame the entire visual content of a room. Inspired by the sRD-SIFT we develop a computational efficient feature detection and matching strategy that substantially benefits the recognition based in visual words. *We also compare the classical BOV against the recent framework of Geometric Preserving Visual Phrases (GVP), showing that the latter outperforms the former.*

3.1 Introduction

One valuable competence for a robot is the ability to localize itself with respect to the environment for performing autonomous navigation [57] and obstacle avoidance [58]. Visual recognition has been used for localization purposes by establishing correspondences between a query image and a database of geo-referenced images constituting

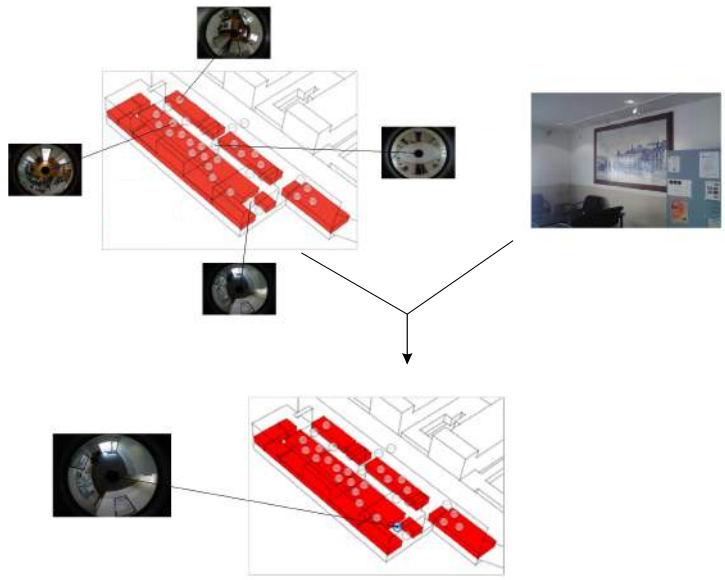


Figure 3.1: Indoor localization scheme using omnidirectional visual maps.

a topological visual map [59]. However, this approach has several difficulties: (i) The query image and the corresponding image in the database, although representing the same visual contents, can substantially differ in appearance (e.g. different lightning, substantial change in viewpoint, etc); (ii) Environments containing symmetric and/or repetitive structures, e.g. doors, walls or corridors, suffer from substantial perceptual aliasing [60]; and (iii) Building a database of large scale environments can be troublesome, specially if we want an exhaustive visual coverage of the environment [59].

Image-based localization based on distinguishable scene landmarks is closely related to image retrieval [3], object recognition [61], and location recognition [59] problems. A commonly adopted scheme extracts local image features [3], quantizes their descriptors to visual words, and applies methods adapted from text search engines to accomplish visual recognition [61, 62]. Many authors take advantage of these techniques, primarily designed for perspective images, for performing image-based localization using omnidirectional images [63]. Typically the image description is accomplished by the extraction of local [3] or global [63] features for topological and metric localization using omnidirectional images in a hierarchical recognition framework [61]. In these prior works, the recognition concerns images acquired us-

ing the same type of imaging system, i.e. perspective [61, 62], or omnidirectional cameras [63].

A closely related work to ours is the one of Chen *et al.* [64], where the authors perform the coverage of a city-scale outdoor environment using a panoramic camera. The authors discuss that performing matching between a perspective query and a database of omnidirectional panoramas leads to poor performance, and propose a rectification process to solve this problem. Instead of using signal reconstruction techniques, which are often subject to interpolation artifacts, we solve the problem by accounting with the distortion during keypoint detection and description. For retrieving the location of the query images we compare two approaches: the classic *bags-of-words* and the recent concept of *visual phrases* [65]. The main difference is that the *visual phrases* introduce weak spatial constraints during the recognition process, while in the standard *bags-of-words* framework the spatial layout of the features is lost.

In this chapter the goal is to perform image-based localization when the query and database images are acquired using different imaging systems (hybrid imaging systems). Taking advantage of the omnidirectional images to perform a complete coverage of the environment, we want to retrieve the location of a query image taken from a conventional camera, e.g. a mobile robot equipped with a perspective camera, or a cell-phone image taken from a person who wants to retrieve its location. While the omnidirectional images permit to speed up the acquisition of thorough visual maps, they also introduce non-linear image distortion that increases the appearance difference between the images. Inspired in the RD-SIFT framework, we propose a keypoint detector and descriptor for omnidirectional images that mitigate this effect, and substancially improves the localization performance.

3.1.1 Chapter Overview

This chapter starts by briefly reviewing the para-catadioptric image formation process [44, 66], and strategies for matching in hybrid imaging systems [32, 67]. Section 3.3 proposes a new framework for feature detection and matching between perspectives and para-catadioptric images, and compares its performance against commonly used strategies for matching in hybrid imaging systems. Section 3.4 evaluates the proposed

method in image-based indoor localization with a database of more than 100 images indexed by 450 perspective queries images.

3.2 Background

3.2.1 Image Formation Model

Barreto and Araujo [44, 66] show that the mapping between points in the 3D world and points in the para-catadioptric image plane can be divided in three steps:

1. Visible points in the scene $\mathbf{X_h}$ are mapped into projective rays/points $\hat{\mathbf{x}}$ in the catadioptric system reference frame that is centered in the effective view point. The transformation is linear and can be described by a 3 x 4 matrix P such that

$$\hat{\mathbf{x}} = \mathsf{P}\mathbf{X_h} = \mathsf{R_c}\,[\,\mathsf{I}\,|\,-\,\mathsf{C}\,]\mathbf{X_h} \tag{3.1}$$

   where C represents the world origin coordinates in the catadioptric system reference frame, $\mathsf{R_c}$ is the rotation matrix between the two coordinate systems, and I is a 3 x 3 identity matrix.

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

$$\bar{\mathbf{x}} = \mathbf{h}(\hat{\mathbf{x}}) = \left(\hat{x} \quad \hat{y} \quad \hat{z} + \sqrt{\hat{x}^2 + \hat{y}^2 + \hat{z}^2}\right)^{\mathsf{T}} \tag{3.2}$$

3. Projective points $\mathbf{x}$ in the catadioptric image plane are obtained after the projective transformation

$$\mathbf{x} = \underbrace{\mathsf{K_c}\begin{bmatrix} 2p & 0 & 0 \\ 0 & 2p & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{\mathsf{H_c}}\bar{\mathbf{x}} \tag{3.3}$$

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

(a) Original image (b) Cylinder

Figure 3.2: Cylindrical panorama obtained from the warping of para-catadioptric image of Figure 3.2(a).

3.2.2 Cylindrical Coordinates

It is possible to obtain virtual perspectives by back-projecting the omnidirectional images into planes. However, we aim at using panoramic images for recognition purposes, making use of the thorough coverage of the environment captured by a single image. For further considerations on how to obtain virtual camera perspectives we point the readers to [44]. It is also possible to map the original image into a cylinder and unfold it to obtain a panorama. Let $\hat{\mathbf{x}}$ be the backprojection of the image point $\mathbf{x}$:

$$\hat{\mathbf{x}} = (\hat{x}, \hat{y}, \hat{z})^T = \mathbf{h}^{-1}(\mathsf{H}_c^{-1}\,\mathbf{x}) \tag{3.4}$$

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

$$\begin{cases} \theta = s \cdot \arctan\left(\dfrac{\hat{x}}{\hat{y}}\right) \\ h = s \cdot \dfrac{\hat{z}}{\sqrt{\hat{x}^2 + \hat{y}^2}} \end{cases} \tag{3.5}$$

with s being a scaling factor (the *radius* of the cylinder). We consider $s = f$, where f is the focal length, in order to minimize the deformation near the center of the image [68]. Figure 3.2(b) is the result of rectifying the para-catadioptric image of Fig. 3.2(a)[1].

[1]The transformation of the catadioptric image to the cylindrical panorama requires the calibration matrix H_c that we obtain using the *CatPack* toolbox made available by Barreto [69]

3.2.3 Matching in Hybrid Imaging Systems

One possible approach to obtain matches between images coming from central catadioptric systems and conventional cameras was proposed by Luis Puig *et al.* [67]. The omnidirectional images are warped using a transformation to polar coordinates using (3.7). SIFT features are computed on the warped and perspective images for establishing putative matches.

$$\theta = \arctan\left(\frac{y}{x}\right) \tag{3.6}$$

$$\rho = \sqrt{x^2 + y^2} \tag{3.7}$$

The generated polar images are very similar to the ones obtained using the mapping to cylindrical coordinates of section 3.2.2. However, the transformation from cartesian to polar coordinates has the advantage of not requiring camera calibration.

Other possible solutions for hybrid matching are the framework describe in section 2.1 that map the into the sphere, like the pSIFT [32] and LB operator [39, 40]. As discussed earlier, such representation minors the problems inherent to planar perspective projection, enabling non-linear distortion invariance and extra invariance to rotation. However, the approach requires perfect camera calibration for both perspective and catadioptric images. In this work, we assume that the perspective camera is not calibrated such that the query images can be acquired by a hand-held device, e.g. cell-phone camera which precludes the usage of the methods that assume the sphere as the underlying image domain.

3.3 Feature Detection and Matching in Hybrid Imaging Systems

This section proposes a new method for extracting image features from omnidirectional images that can be reliable matched with perspective image features. Instead of rectifying the omnidirectional image to perspective images [51], we implicitly compensate the distortion effect based on the rectification to cylindrical coordinates, which

enables the use of the wide field-of-view of the omnidirectional images. Finally, we evaluate the proposed method using standard repeatability and precision-recall tests, and compare it against some approaches for matching between mixtures of perspective and para-catadioptric images.

3.3.1 SIFT for Cylindrical Images

Keypoint detection

The objective here is to generate a scale-space representation equivalent to the one that would be obtained by filtering the cylindrical panorama. Instead of explicitly computing a new image using signal reconstruction techniques, which are often subject to interpolation artifacts [50, 51], we adapt the convolution kernels to directly process the para-catadioptric image samples.

Through the manipulation of Eq. 3.4 and Eq. 3.5, we can re-write the mapping from para-catadioptric coordinates to cylindrical coordinates as

$$\mathbf{u} = \mathbf{\Psi}(\mathbf{x}) = \begin{pmatrix} \Psi_u(x,y) \\ \Psi_v(x,y) \end{pmatrix} = \begin{pmatrix} f \cdot \arctan(x/y) \\ \dfrac{f^2 - r^2}{2r} \end{pmatrix}. \tag{3.8}$$

The inverse of Eq. 3.8 provides the mapping between cylindrical and para-catadioptric coordinates:

$$\mathbf{x} = \mathbf{\Psi}^{-1}(\mathbf{u}) = \begin{pmatrix} \Psi_x^{-1}(u,v) \\ \Psi_y^{-1}(u,v) \end{pmatrix} = \begin{pmatrix} y \tan(\dfrac{u}{f}) \\ cos\left(\frac{u}{f}\right) \left(\sqrt{f^2 + v^2} - v\right) \end{pmatrix}. \tag{3.9}$$

Let's now consider the convolution of the cylindrical image $\mathtt{I}^{cyl}$ with a Gaussian kernel with standard deviation σ. By writing the convolution operation of Eq. 2.1 explicitly, it comes that the blurred image is

$$\mathtt{L}_\sigma^{cyl}(s,t) \;=\; \sum_u \sum_v \; \mathtt{I}^{cyl}(u,v)\, \mathsf{G}_\sigma(s-u, t-v)\,. \tag{3.10}$$

Following the same reasoning of section 2.4.2, it is possible to obtain the adaptive

filtering of Eq. 3.11

$$\mathtt{L}_\sigma(h,k) \;=\; \sum_x \sum_y \; \mathtt{I}(x,y)\, \mathtt{G}_\sigma\Big(f{\cdot}\Big(\arctan\Big(\frac{h}{k}\Big){-}\arctan\Big(\frac{x}{y}\Big)\Big)\,,\; \frac{f^2(\delta-1)+\delta r^2(\delta-1)}{2\delta r}\Big), \tag{3.11}$$

with r being the distance between the center and the image location where the filter is applied

$$r \;=\; \sqrt{h^2+k^2}\,, \tag{3.12}$$

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

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

Note that now the smoothing convolution is an operation of $\mathbb{R}^2 \times \mathbb{R}^4 \rightarrow \mathbb{R}_+$ due to its dependence in (h,k) and (x,y). For each radius, the adaptive blurring kernel has the same shape, but with different orientations (see Fig.3.3). Like the RD-SIFT, the pixel shape and orientation depends on its position in the image, becoming computationally heavy to filter the image. To leverage such computational burden, we approximate the adaptive filters by a rank 1 filter that can be written as the outer product of two 1D Gaussian filters of the same standard deviation. This permits to implement the convolution process separately in X and Y dimensions like in the sRD-SIFT algorithm, which permits to considerably speed up the smoothing process [34]. Instead of computing the *cylindrical* Gaussian for each image pixel position, we approximate (3.11) by the closest rank 1 Gaussian filter estimated using Singular Value Decomposition

$$\Big[\mathsf{U} \quad \mathsf{S} \quad \mathsf{V}\Big] = \mathtt{SVD}(\mathsf{G}_\sigma). \tag{3.13}$$

Thus, the rank 1 Gaussian kernel that better approximates G_σ is

$$\mathsf{G}_{\sigma,rank=1} = \mathsf{U}_{(:,1)}\mathsf{S}_{(1,1)}{\mathsf{V}_{(:,1)}}^{\mathsf{T}} \tag{3.14}$$

where $\mathsf{S}_{(1,1)}$ representing the first singular value and $\mathsf{V}_{(:,1)}$ representing the correspondent singular vector (see Fig.3.3 for an illustration of the process). We have observed

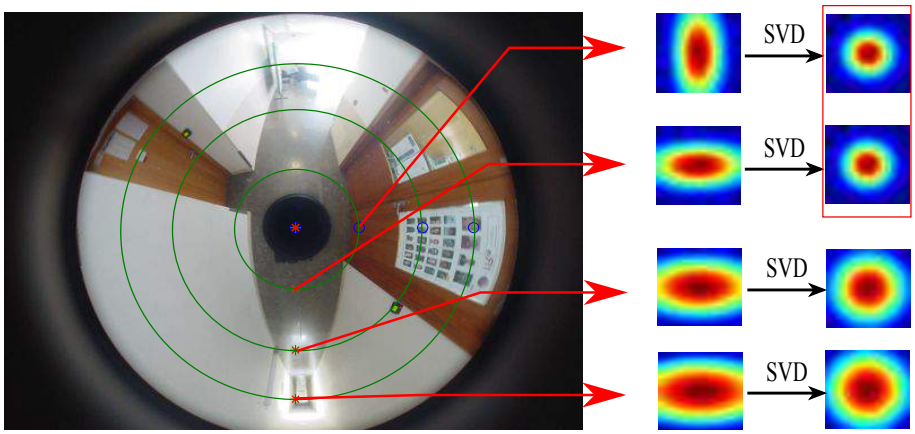


Figure 3.3: Separable filters for catadioptric images.

that this decomposition has two significant advantages: (i) For every image radius the same $\mathsf{G}_{\sigma,rank=1}$ can be used, which enables separable convolution for each radius in a similar way to [51]; and (ii) a filter bank can be computed offline and loaded into memory when required. We consider the same filter bank for all the para-catadioptric images used throughout this chapter.

Keypoint description

Instead of explicitly correct the local image patches for the SIFT descriptor computation, we use the same reasoning of section 2.5 to implicitly correct the gradient by measuring the gradients in the original image and correct the result using the derivative chain rule. Let $\mathtt{I}$ be the catadioptric image and $\mathtt{I}^{cyl}$ be the cylindrical panorama. The mapping relation between the two images is the following:

$$\mathtt{I}^{cyl}(\mathbf{u}) \;=\; \mathtt{I}(\boldsymbol{\Psi}^{-1}(\mathbf{u}))\,.$$

Applying the derivative chain rule it yields

$$\nabla\mathtt{I}^{cyl} \;=\; \mathsf{J}_{\boldsymbol{\Psi}^{-1}}\,.\,\nabla\mathtt{I} \tag{3.15}$$

with $\nabla\mathtt{I}^{cyl}$ and $\nabla\mathtt{I}$ being respectively the gradient vectors in $\mathtt{I}^{cyl}$ and $\mathtt{I}$, and $\mathsf{J}_{\boldsymbol{\Psi}^{-1}}$ being the 2×2 Jacobian matrix of the mapping relation given in (3.9). The Jacobian

(a) Set A (b) Set B (c) Set C (d) Set D

Figure 3.4: Example of the data sets used for detection and description evaluation.

matrix can be written in terms of para-catadioptric image coordinate $\mathbf{x} = (x, y)^\mathsf{T}$:

$$\mathsf{J}_{\mathbf{\Psi}^{-1}} = \begin{pmatrix} \frac{r^2}{fy} & 0 \\ -\frac{x}{fr}\left(\tau + \sqrt{\tau^2 + f^2}\right) & \frac{y}{r}\left(\frac{\tau+\sqrt{f^2+\tau^2}}{\sqrt{f^2+\tau^2}}\right) \end{pmatrix},$$

with $\tau = \frac{r^2-f^2}{2r}$. The final descriptor is generated from the undistorted gradients $\nabla \mathtt{I}^{cyl}$ following the procedure described in section 2.4.1. This framework for keypoint detection and matching is called SIFT for Cylindrical Images (cylSIFT).

3.3.2 Performance Evaluation in Planar Textured Surfaces

Methods under evaluation

In this hybrid matching comparison, SIFT [3] is always used to extract features in the perspective images and the test only differ in terms of the method used to extract features in the para-catadioptric/rectified views. We compare the proposed cylSIFT method against the following approaches: Application of SIFT over (i) para-catadioptric images (SIFT); (ii) rectification to polar coordinates (*Polar*); (iii) rectification to cylindrical coordinates (*Cylinder*); and (iv) Virtual Camera Perspectives (VCP). To generate the VCP we manually select the region in the omnidirectional images that correspond to the visual contents of the perspectives. Without this prior knowledge we would need to render 4 or more perspectives for each omnidirectional image, and still be subject to viewpoint changes arising in the synthetically generated perspective images. Although the VCP is not a direct competitor of our method because it does not encapsulate the same wide field-of-view in one image, it is the theoretical top

performer since matching is accomplished between images with no distortion, being included in the performance evaluation study for the sake of completeness.

Metrics for evaluation

In terms of detection evaluation, the repeatability of keypoint detection is unarguably the most important property of a reliable detector [45]. Let's consider S_{cata} and S_{pers} as being the set of keypoints detected in the para-catadioptric image (or rectifications obtained from it) and perspective images, respectively. Given two images of the same scene, the repeatability measures the percentage of the features detected on the scene part visible in both images:

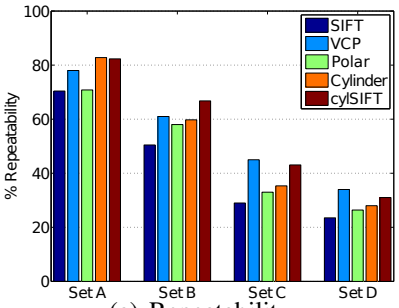
$$\%_{\text{Repeatability}} = \frac{\#(S_{cata} \cap S_{pers})}{\#S_{pers}} * 100 \tag{3.16}$$

where $\#$ denote the cardinality of the sets. For matching evaluation we use the traditional 1-precision *vs* recall curves [46] early introduced in section 2.3.2.

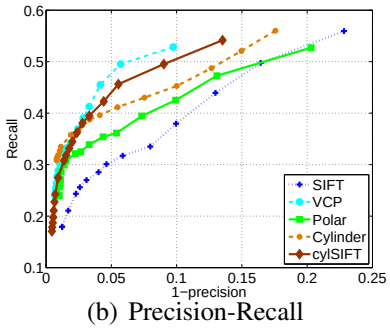
Datasets

We collected 13 para-catadioptric images taken in different places using a camera with a resolution of 2272×1704. On the perspective side we collect 4 different perspective image sets (Fig. 3.4): set (A) was acquired fronto-parallel to the scene, at the same location of the para-catadioptric system; set (B) was acquired from the same position as the para-catadioptric image and with an angle of approximately 45 degrees between the optical axis and the vertical plane; set (C) presents strong scale changes while preserving the fronto-parallel viewpoint; and set (D) was taken from different positions and viewpoints relatively to the para-catadioptric images, to test strong viewpoint changes. The resolution of the perspective images is 1600x1200. Similarly to the evaluation with planar textured surfaces of the previous chapter, at this stage we only consider images of planar scenes that enables to find a ground truth homography[2] for verification of detection and matching results,

[2]The ground truth homography is computed after rectifying the para-catadioptric coordinates to perspective coordinates.



(a) Repeatability



(b) Precision-Recall

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

Figure 3.5: Detection and description evaluation in planar image pairs. Fig. 3.5(a) compares the repeatability scores of the several methods evaluated, while Fig. 3.5(b) concerns description evaluation. We can observe that using the cylSIFT approach permits to have similar scores to the rectification for a perspective view. Additionally, we provide the average running time of every method, number of detection and number of matches established using the similarity distance thresholded at 0.9. The computation differences between the SIFT and the cylSIFT rely on the offline computation of the filter bank, which in our Matlab implementation takes in average 1 second, and in the gradient correction technique. In VCP, Polar and Cylinder the rectification process using our Matlab routines is included. $(\cdot)^{+}$ denotes that for the VCP we only show the running the time for the correct perspective. In practice at least 4 perspective images must be rendered for each omnidirectional to cover its wide field of view.

Results and discussion

The repeatability of detection and precision-recall curves for description can be observed in Fig. 3.5. We can observe that the cylSIFT performs better than most competing methods over the panoramic images. The image re-sampling for distortion compensation requires the reconstruction of the discrete image signal. This reconstruction process can either remove high frequency components and/or introduce new spurious frequencies being highly prejudicial in the detection step [51]. The Polar and the Cylinder generate similar images and it is expected that both provide similar results. However, as the rectification to cylindrical coordinates uses the calibration

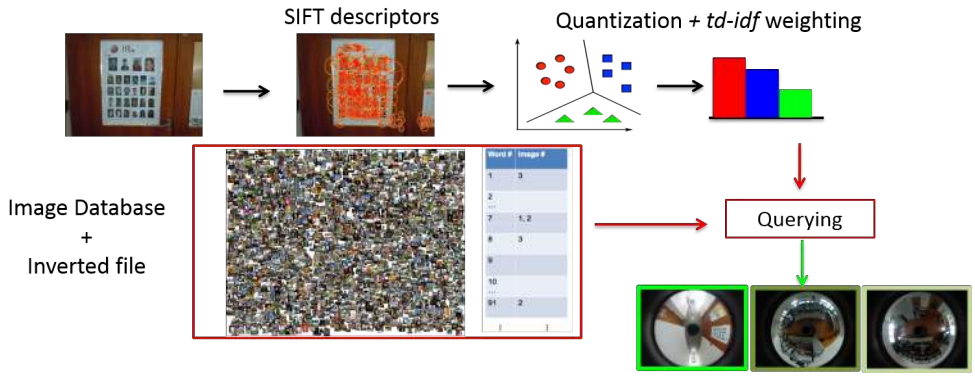


Figure 3.6: Indoor image-based localization pipeline.

matrix and the non-linear function characteristic of the mirror, the mapping of the latter is more accurate than the former, which explains the observed better performance. The cylSIFT method performs very closely to the VCP approach, showing that, even dealing with the distortion on the cylinder, the cylSIFT is capable of performing close the perspectives generated through interpolation.

In terms of description, we can observe that performing implicit gradient correction provides gains in terms of matching performance, when compared with the other descriptors computed in the panoramic images. Once more it is verified that the VCP provides the best matching scores, which is expected since the description space, although subject to interpolation artifacts, does not present any non-linear distortion.

In summary, we can conclude that the cylSIFT outperforms SIFT applied directly over the omnidirectional image, as well as polar and cylindrical panoramas. The VCP approach outperforms the cylSIFT algorithm due to the correct alignment between the perspective image and generated VCP (see results of set D in Fig. 3.5(a)). In a real application scenario, this correct alignment is not known in advance, precluding a good performance for this method.

3.4 Indoor Localization with Hybrid Imaging Systems

In this section we evaluate the proposed cylSIFT method for image-based localization. Given a query image, acquired with a standard camera (e.g. robot or a person with a

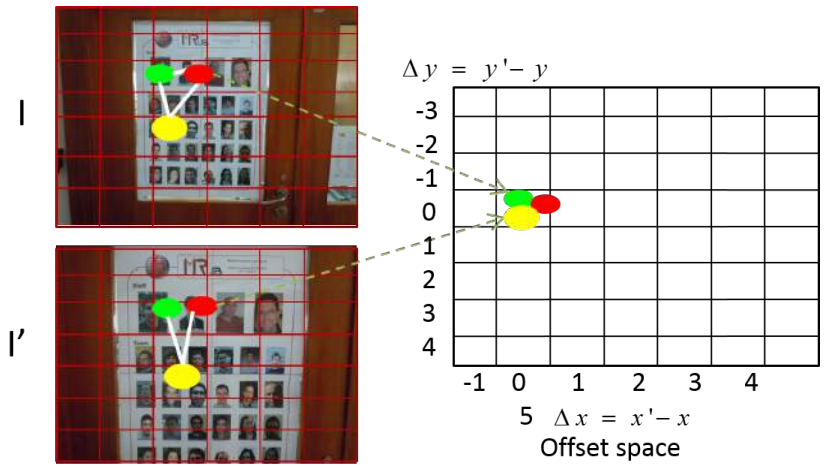


Figure 3.7: Illustration of a GVP occurrence. The process starts by quantizing the image space into an offset grid. The spatial layout of the visual words are then used to build GVP occurrences.

conventional camera), the localization is obtained by searching and retrieving the most similar view in a database of omnidirectional visual maps. For a visual illustration of the process see Fig. 3.6.

3.4.1 Retrieval Schemes

In our retrieval application, we compare two different searching approaches. The first method uses the standard BOV approach. A vocabulary tree is built using k-means clustering. The basic idea of descriptor clustering is to represent similar local image descriptor with the same visual words, which results in very high dimensionality reduction and subsequent speed-up when querying large image databases. Since the SIFT descriptor is 128D, the process of clustering is speed -up by using hierarchical k-means where each branch is recursively spitted into k new groups along L-levels of the tree, which totalizes k^L visual words. The correspondence between images is given by measuring the similarity between the visual words in a query image and in the database images [61]. Although this scheme provides good performance in several recognition scenarios [61, 63], it discards the spatial relation of the visual words during retrieval that can be relevant to disambiguate situations of perceptual aliasing [59].

The second method uses the new concept of *visual phrases* (*GVP*) [65]. The objective of using GVP is to take into account the spatial relations between visual words. For each pair of the same word in the query and database images, the offset is computed by subtracting their corresponding locations. A set of n visual words in a certain spatial layout define a GVP of length n. The image space is quantized into cells to tolerate shape deformation and to build an efficient voting scheme. After computing the offset, a vote is generated on the offset space. n votes in the same offset cell correspond to a co-occurring GVP of length n. Refer to Fig. 3.7. for the illustration of a GVP occurrence.

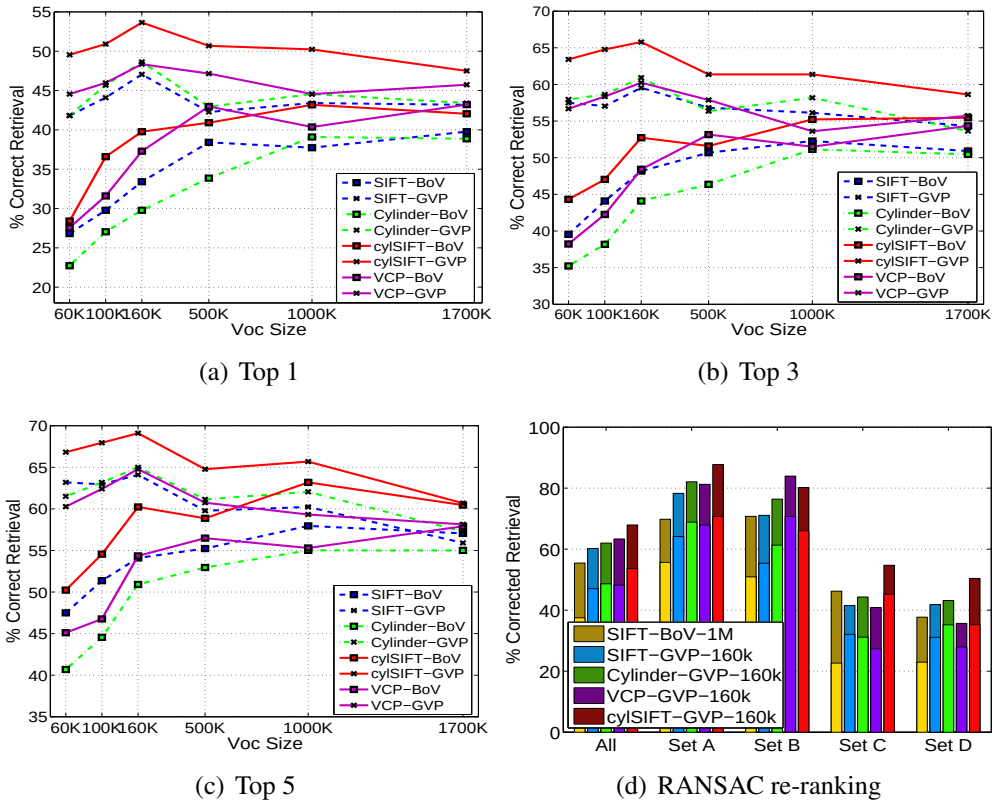
3.4.2 Feature Extraction Methods and Database considerations

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

The feature extraction techniques and searching schemes are tested by performing queries on a database of 118 para-catadioptric images that provide a detailed visual map of the building where the robot operates. Concerning the VCP database, we render 4 perspectives for each omnidirectional image. Each generated perspective image has a field of view of 108° and resolution of 1600×1200. Unlike in the tests of section 3.3, the VCP images are generated in an unsupervised manner, meaning that each omnidirectional image gives raise to 4 VCP without assurance that one of the VCP is aligned with the perspective query image. We use 451 query images for evaluating which combination of retrieval scheme, vocabulary size and feature extraction method (at the database side) performs better for the task.

The performance of retrieval is given by the percentage of correctly retrieved locations in first place (Top 1), and in the sets of 3 and 5 images with highest scores (Top 3 and Top 5). Finally, the best retrieval method for each feature extraction technique is selected and the top 5 images are re-ranked through geometrical verification within a RANSAC framework [21].

3.4.3 Results and Discussion



(a) Top 1 (b) Top 3

(c) Top 5 (d) RANSAC re-ranking

Figure 3.8: Retrieval results in indoor environment. We have tested several combinations regarding feature extraction techniques, vocabulary size and searching scheme. The cylSIFT method proved to be superior to the other feature extraction approaches, regardless of the type of retrieval scheme and vocabulary size. For each feature extraction method, we selected the best retrieval scheme and performed re-ranking on the top 5 images using strong geometric constraints within a RANSAC framework (Fig. 3.8(d)).The darker colors represent the improvement obtained over the GVP framework, with a vocabulary size of 160k. We additionally include the scores of a naive approach to the problem where SIFT features and standard BOV are used for localization recognition.

Figure 3.8 presents the retrieval results. The cylSIFT approach is the one providing the highest retrieval scores, independently of the searching scheme and vocabulary size. Increasing the number of words in the vocabulary increases the performance of the BOV approach. In this case, the recognition is performed in a word-by-word

basis and more discriminative words tend to provide better retrieval results. We also observed that using a lower vocabulary size tends to favor the performance of the GVP. Since vocabularies of small sizes are less discriminative more common words between the query and the database image can be established and more visual phrases exist, leading to a boost in performance.

For each feature extraction method, we selected the best retrieval scheme (GVP with a vocabulary size of $160k$) and performed re-ranking on the top 5 images using strong geometric constraints within a RANSAC framework (Fig. 3.8(d)). In addition to the average correct retrieval scores, we also provide the results for each perspective sets. The higher quality of matching provided by our method can be clearly seen for all the 4 sets, but with particular emphasis in the most difficult ones (set C and D). It is also important to notice that in set D the VCP tends to be outperformed by all other methods. This is due to the fact that in general the perspective image is misaligned with the generated VCP, meaning that such schemes is only effective if we known in advance which region of the omnidirectional image is being viewed to ensure minimal viewpoint changes.

One important observation is that the interpolation used in the explicit cylindrical and VCP images has a negative impact in the visual words distinctiveness. Using the BOV with descriptors extracted in cylindrical images does not lead to an increase in performance when comparing with SIFT, showing that the visual words computed on these descriptors are less discriminative. While in the section 3.3, two descriptors were considered a match by using similarity distance (nearest neighbor distance ratio) [46], in the vocabulary tree, two descriptors belong to the same visual word if they are close to the same centroid. Therefore, the smaller the euclidean distance between two descriptors, the greater the probability of belonging to the same visual word. Although the interpolation artifacts do not have a large influence in the nearest neighbor ratio, they seem to be particularly relevant for the computation of the image visual words. The implicit filtering approach seems to be immune to this phenomena and takes full advantage of its higher matching performance.

3.5 Closure

This chapter focus on indoor image-based localization by querying omnidirectional maps using perspectives. We take advantage of the wide field-of-view of the images, which enable a complete description of the environment with minimum effort. To successfully retrieve the omnidirectional image using a perspective, we develop a new algorithm for feature detection and description based on the rectification to cylindrical images. Extensive experiments prove that our method outperforms explicit image rectification methods, proving to be beneficial for image-based localization by improving the rate success rate in 15%.

Chapter 4

Image Alignment in the presence of Radial Distortion

Image alignment consists in finding the deformation between a reference and an incoming image through the minimization of an intensity-based cost function. Solving this problem typically involves the assumption of an image motion model (warping function) that describes the expected deformation a reference template suffers between two time instants. In this chapter, we study the problem of image alignment in wide FOV cameras, and we propose a set of motion models that implicitly encompass the distortion effect arising in this type of imaging devices. We show that including the proposed motion models in a inverse compositional alignment framework enables to recover the image radial distortion whenever the camera is not calibrated.

4.1 Introduction

Image alignment consists in finding the deformation between a template and an incoming image through the minimization of an intensity-based cost function. Since the seminal work of Lucas and Kanade [70] on optical flow, image alignment has been applied in a broad range of applications, such as tracking [71], medical image registration [72], and face alignment [73]. In the past two decades, several authors

devoted attention to the original method of [70] by improving optical flow accuracy in wide baseline situations [74–77], or by manipulating the motion models for increased robustness against illumination changes [78].

Image alignment techniques often rely on the assumption of a motion model that describes the deformation expected between the template and the incoming images. Several motion models are currently used, ranging from a low complexity translation model [70, 79] to an affine motion model [75, 80, 81]. Unfortunately, these motion models do not compensate the RD effect arising in cameras equipped with unconventional optics. At the image level, the distortion causes a non-uniform displacement of the pixel positions along radial directions and towards the center, which introduces a non-linear image deformation that conventional image models do not tolerate. In practical terms, the inability of the standard motion models for accommodating RD translates in localization drifts and, more importantly, affect the registration accuracy [82, 83].

Despite these facts, image alignment has been applied in the past to images with significant RD [84, 85], mainly in the context of feature tracking (Lucas-Kanade-Tommasi tracker (KLT)). Some works directly apply the KLT method over RD images and, therefore, violate the underlying assumptions of the KLT tracker, which were done for perspective images. Other solutions used in the literature either discard the image boundaries [84], where the distortion effect is more pronounced, or correct the distortion in a pre-processing step before applying the KLT. Although the later approach is quite straightforward, the distortion rectification requires the interpolation of the image signal, which can be computationally expensive, and, even more important, unreliable since the synthetically corrected images contain artificially interpolated pixel intensities [29, 51].

This chapter focus on image alignment in images presenting significant radial distortion. We propose an extension of the standard perspective motion model for describing the image template deformation that fuses local motion with global image distortion. Unfortunately, the particular structure of this warp does not allow to calibrate the distortion during tracking, as it will be theoretical explained latter. To cope with this problem, we additionally propose an approximation to the ideal theoretical

model that enables to calibrate distortion during tracking. To the best of our knowledge, this is the first work showing that is possible to estimate RD through the registration of image patches. Photometric deformations will not be considered for clarity. Affine photometric models [78] can be introduced without changing the underlying results.

4.1.1 Related Work

Optical flow computation in cameras equipped with unconventional optics was, probably for the first time, discussed by Daniilidis *et al.* [29] that show that optical flow on catadioptric images should be computed assuming the sphere as underlying domain of the image function to deal with the non-uniform sampling of catadioptric images.

Mei *et al.* propose in [86, 87] a region tracking algorithm for generic central cameras where the warping is also formulated on the sphere. The approach is specific to the tracking of plane surfaces and requires the camera to be calibrated. In [88], Salazar *et al.* use the warping function proposed in [86, 87] to perform homography-based tracking in uncalibrated images by simply adding the camera intrinsics to the vector of unknown parameters to be estimated. The work of Salazar *et al.* is still specific to the tracking of large plane surfaces, it involves computationally expensive minimization that precludes real-time performance, and it requires tracking across three or more frames to recover the camera parameters [88].

A closely related work to ours is the one of Tamaki *et al.* [89] that propose an image alignment approach to calibrate the camera radial distortion. The method registers a distortion-free planar pattern with a distorted view of this pattern, and uses non-linear optimization to estimate the plane homography under perspective, the radial distortion, and the linear spatial changes in illumination. Like our method, the algorithm just requires two views for computing the warping parameters, but the requirement of a distortion-free view of the pattern limits usability.

Despite of being less general than [86–88] in the sense that it can only be applied to cameras where the division model is valid, our method does not require the camera intrinsic calibration to be known [86, 87], and it is able to recover the real radial distortion parameter solely by tracking low-level features between adjacent frames. Our

method shares some similarities with [89], but does not require a calibration pattern, and recovers distortion from the motion of low-level image patches.

4.1.2 Chapter Overview

The structure of this chapter is as follows: Section 4.2 reviews the literature related with image alignment. Section 4.3 derives the RD compensated motion models and explains how to include them in the inverse compositional alignment framework. Special focus is given to the case of uncalibrated images, where a computational efficient method that enables to simultaneously estimate global distortion and local feature motion is presented. The performance of the proposed motion models is first validated in a synthetic data set, where we study the effect of number and size of the templates in the estimation of distortion. In section 4.5, the proposed motion models are evaluated in feature tracking applications with a representative set of repeatability [82] and medical endoscopy SfM experiments [51].

4.2 Background

Along this chapter it is assumed that the RD can be fairly described using the division model introduced in chapter 2. In this section, we review the image registration frameworks with direct and inverse image alignment. We also summarize standard image motion models, and discuss the importance of the local template updates and pyramidal image representation for achieving reliable long-term template tracking.

4.2.1 Image Alignment Framework

Image alignment between temporally adjacent images can be formulated as a non-linear optimization problem whose cost function is the sum-of-squared differences between a template $\mathtt{T}$ and incoming images $\mathtt{I}$. The goal is to compute

$$\epsilon = \sum_{\mathbf{u} \in \mathcal{N}} \Big[\mathtt{I}(\mathbf{w}(\mathbf{u}; \mathbf{m})) - \mathtt{T}(\mathbf{u}) \Big]^2, \tag{4.1}$$

where $\mathbf{m}$ denotes the components of the image warping (or motion model) function $\mathbf{w}$, and $\mathcal{N}$ denotes the integration region of an image point $\mathbf{u}$. Lucas and Kanade proposed to minimize Eq. 4.1 by assuming that a current motion vector $\mathbf{m}$ is known and iteratively solve for $\delta\mathbf{m}$ increments on the warp parameters, with Eq. 4.1 being approximated by

$$\epsilon = \sum_{\mathbf{u}\in\mathcal{N}} \Big[\mathtt{I}(\mathbf{w}(\mathbf{u};\mathbf{m}+\delta\mathbf{m})) - \mathtt{T}(\mathbf{u})\Big]^2 \approx \sum_{\mathbf{u}\in\mathcal{N}} \Big[\mathtt{I}(\mathbf{w}(\mathbf{u};\mathbf{m})) + \nabla\mathtt{I}\frac{\partial\mathbf{w}}{\partial\mathbf{m}}\delta\mathbf{m} - \mathtt{T}(\mathbf{u})\Big]^2. \tag{4.2}$$

Differentiating ϵ with respect to $\delta\mathbf{m}$, and after some algebraic manipulations, a closed form solution for $\delta\mathbf{m}$ can be obtained:

$$\delta\mathbf{m} = \mathcal{H}^{-1}\sum_{\mathbf{x}\in\mathcal{N}} \Big[\nabla\mathtt{I}\frac{\partial\mathbf{w}(\mathbf{x};\mathbf{m})}{\partial\mathbf{m}}\Big]^{\mathsf{T}} \Big(\mathtt{T}(\mathbf{x}) - \mathtt{I}(\mathbf{w}(\mathbf{x};\mathbf{m}))\Big), \tag{4.3}$$

with $\mathcal{H}$ being a 1st order approximation of the Hessian matrix [80,81], and the parameter vector being additively updated $\mathbf{m}^{i+1} \leftarrow \mathbf{m}^i + \delta\mathbf{m}$ at each iteration i. This method is also known as *forward additive alignment* [80, 81] and it requires to re-compute $\mathcal{H}$ at each iteration due its dependence with $\mathtt{I}$.

For efficiently solving Eq. 4.2, Baker and Matthews [80, 81] proposed an *inverse compositional alignment* method that starts by switching the roles of $\mathtt{T}$ and $\mathtt{I}$

$$\epsilon = \sum_{\mathbf{u}\in\mathcal{N}} \Big[\mathtt{I}(\mathbf{w}(\mathbf{u};\mathbf{m})) - \mathtt{T}(\mathbf{w}(\mathbf{u};\delta\mathbf{m}))\Big]^2 \tag{4.4a}$$

$$\approx \sum_{\mathbf{u}\in\mathcal{N}} \Big[\mathtt{I}(\mathbf{w}(\mathbf{u};\mathbf{m})) - \mathtt{T}(\mathbf{w}(\mathbf{u};\mathbf{0})) - \nabla\mathtt{T}\frac{\partial\mathbf{w}}{\partial\mathbf{m}}\delta\mathbf{m}\Big]^2. \tag{4.4b}$$

The increments $\delta\mathbf{m}$ are then computed as:

$$\delta\mathbf{m} = \mathcal{H}^{-1}\sum_{\mathbf{u}\in\mathcal{N}} \Big[\nabla\mathtt{T}\frac{\partial\mathbf{w}(\mathbf{u};\mathbf{0})}{\partial\mathbf{m}}\Big]^{\mathsf{T}} \Big(\mathtt{I}(\mathbf{w}(\mathbf{u};\mathbf{m})) - \mathtt{T}(\mathbf{u})\Big), \tag{4.5}$$

with $\mathbf{w}(\mathbf{u};\mathbf{0})$ being the identity warp. $\mathcal{H}$ is computed using the template gradients and, therefore, it is constant during the registration procedure, leading to a significant computational improvement when compared with the forward additive alignment. Finally,

the warp parameters are updated as follows:

$$\mathbf{w}(\mathbf{u};\mathbf{m}^{i+1}) \leftarrow \mathbf{w}(\mathbf{u};\mathbf{m}^{i}) \circ \mathbf{w}^{-1}(\mathbf{u};\delta\mathbf{m}), \tag{4.6}$$

where $\circ$ denotes the composition operator. Although the update rule of the inverse compositional alignment is computationally more costly than a simple additive rule, Baker and Matthews [80, 81] show that the overall computational complexity of the inverse formulation is significantly lower than that of the forward additive KLT.

Motion models for perspective images

The motion model (or image warping function) $\mathbf{w}$ used for in the image alignment framework determines the degree of image deformation tolerated during the registration process. The original contribution of Lucas and Kanade [70, 79] assumes that the neighborhood $\mathcal{N}$ around a feature point $\mathbf{u}$ moves uniformly and, therefore, the authors model the image motion using a simple translation model. However, the deformation that it tolerates is not sufficient when the tracked image region is large, or the video sequence undergoes considerable changes in scale, rotation and viewpoint. In these situations, the affine motion model [75, 80] is typically adopted

$$\mathbf{w}(\mathbf{u};\mathbf{m}) = (\mathsf{I} + \mathsf{A})\mathbf{u} + \mathbf{t}, \tag{4.7}$$

where the parameter vector is $\mathbf{m} = (a_1, .., a_4, t_x, t_y)^{\mathsf{T}}$, and I is a 2×2 identity matrix. Although we work specifically with the affine motion model, the extensions proposed can also be done with for other motion models, such as translation, similarity transformations and homographies.

Pyramidal image representation for the iterative minimization

Despite of the warp complexity, the registration process may fail to converge when the initialization of the warp parameters $\mathbf{m}^0$ is not close enough to the current motion parameters, i.e. $\mathbf{m}^0$ is not in the convergence region $\mathcal{C}$ where the 1^{st} order approximation of Eq. 4.4b is valid [81]. This effect can be attenuated by performing track-

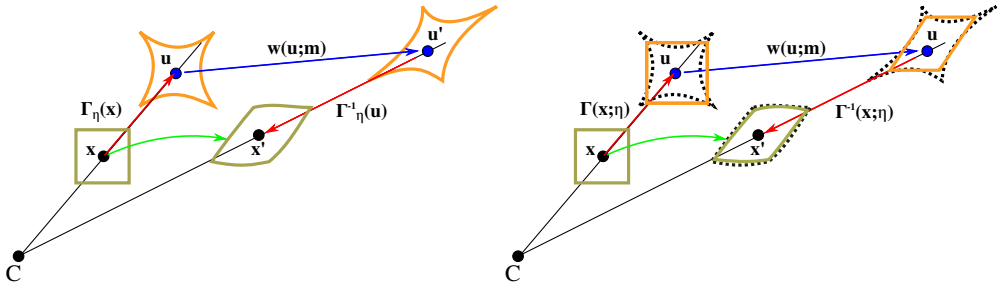
ing using a pyramidal image representation [74], where several image resolutions are built by downsampling the original image signal by factors of 2. A L-levels pyramidal tracking algorithm proceeds from the coarse to the finest pyramid level, with the coarsest feature position being given by $\mathbf{u}^L = 2^{-L}\mathbf{u}$. The registration proceeds at each pyramid level, with the result begin propagated to the next level as $\mathbf{u}^{L-1} = 2\,\mathbf{u}^L$ (for further details see [74]). Since the integration region $\mathcal{N}$ is kept constant across scales, the pyramidal framework greatly improves the probability of $\mathbf{m}^0$ belonging to $\mathcal{C}$, which by consequence increases the tracking success. The number of levels L of the image pyramid representation is typically computed as function of the image resolution [74, 75].

Template update for long-term template tracking

In case of applying alignment techniques in the context of feature (position) tracking in continuous video, it is typically more important to track the position $\mathbf{u}$ than the template itself. Therefore, the template update is a critical step to keep plausible tracks along long image sequences. An inherent problem to the template update step is the localization error introduced whenever the template is updated [83]. High-order motion models tend to be more flexible in terms of the deformation tolerated during the registration process, with the templates being updated less frequently. This minimizes the drift in the feature localization introduced whenever a new template is captured [81, 83]. When applying the proposed motion models in feature tracking applications, a new template is captured whenever the squared error of Eq. 4.1 falls above some threshold [75].

4.3 Image Alignment in Images with Radial Distortion

In this section, we propose extensions of the standard perspective motion models for cameras equipped with lens that introduce significant radial distortion. We address the image alignment problem both in calibrated and uncalibrated cameras, and we show how the distortion can be efficiently estimated for when several templates are being tracked/registered.



(a) $\mathbf{v}_\eta$ (b) $\mathbf{v_p}$

Figure 4.1: Schematic difference between the (a) accurate and the (b) approximate RD compensated motion model. The black dashed lines in (b) represent the patches using the accurate RD model.

4.3.1 Radial Distortion compensated Motion Model

Let's consider the standard situation where we aim at aligning two undistorted images $\mathtt{I}^u$ and $\mathtt{I}^{u'}$ that are related by a generic motion function $\mathbf{w}$, such that $\mathtt{I}^u(\mathbf{u}) = \mathtt{I}^{u'}(\mathbf{w}(\mathbf{u};\mathbf{m}))$. We now consider that $\mathtt{I}^u$ and $\mathtt{I}^{u'}$ are result of removing the radial distortion from $\mathtt{I}$ and $\mathtt{I}'$, respectively. Using the distortion function of Eq. 2.4, we know that corresponding undistorted and distorted coordinates are related by $\mathbf{u} = \mathbf{\Gamma}_\eta(\mathbf{x})$, so we can re-write the mapping relation as $\mathtt{I}^u(\mathbf{u}) = \mathtt{I}^{u'}(\mathbf{w}(\mathbf{\Gamma}_\eta(\mathbf{x});\mathbf{m}))$. Since $\mathtt{I}^u(\mathbf{u}) = \mathtt{I}(\mathbf{x})$, with $\mathbf{x} = \mathbf{\Gamma}_\eta^{-1}(\mathbf{u})$, we can write the mapping that relates two distorted image signals as $\mathtt{I}(\mathbf{x}) = \mathtt{I}'(\mathbf{\Gamma}_\eta^{-1}(\mathbf{w}(\mathbf{\Gamma}_\eta(\mathbf{x});\mathbf{m})))$. The RD compensated motion model that directly relate two distorted image signals can be expressed using the following function composition:

$$\mathbf{x}' = \mathbf{v}_\eta(\mathbf{x};\mathbf{m}) = \left(\mathbf{\Gamma}_\eta^{-1} \circ \mathbf{w} \circ \mathbf{\Gamma}_\eta\right)(\mathbf{x};\mathbf{m}). \tag{4.8}$$

Intuitively, this warping function $\mathbf{v}_\eta$ encompasses three steps: (i) compensates the radial distortion, (ii) applies the motion model, and (iii) restores the non-linear image deformation. In the case of an ideal model, the perspective motion model $\mathbf{w}$ parameters will be the same as the images would be free of distortion. This motion model can be included with minimal effort in the inverse compositional alignment framework.

4.3.2 Image Alignment in Calibrated Images

In case the camera is calibrated and η is known in advance, the parameter vector $\mathbf{m}$ of $\mathbf{v}_\eta$ comprises the same parameters of the original motion of Eq. 4.7. By replacing our motion model $\mathbf{v}_\eta$ in the inverse compositional alignment, it is straightforward to obtain the closed-form solution for $\delta\mathbf{m}$, which is given by:

$$\delta\mathbf{m} = \mathcal{H}_d^{-1} \sum_{\mathbf{x}\in\mathcal{N}} \Big[\nabla \mathtt{T} \frac{\partial \mathbf{v}_\eta(\mathbf{x};\mathbf{0})}{\partial \mathbf{m}}\Big]^\mathsf{T} \Big(\mathtt{I}(\mathbf{v}_\eta(\mathbf{x};\mathbf{m})) - \mathtt{T}(\mathbf{x})\Big) \tag{4.9}$$

with the 1^{st} approximation of the hessian being

$$\mathcal{H}_d = \sum_{\mathbf{x}\in\mathcal{N}} \Big[\nabla \mathtt{T} \frac{\partial \mathbf{v}_\eta(\mathbf{x};\mathbf{0})}{\partial \mathbf{m}}\Big]^\mathsf{T} \Big[\nabla \mathtt{T} \frac{\partial \mathbf{v}_\eta(\mathbf{x};\mathbf{0})}{\partial \mathbf{m}}\Big],$$

and the Jacobian $\frac{\partial \mathbf{v}_\eta(\mathbf{x};\mathbf{0})}{\partial \mathbf{m}}$ being evaluated at $\mathbf{m} = \mathbf{0}$. Finally, the motion parameters are updated at each iteration as follows:

$$\mathbf{v}_\eta(\mathbf{x};\mathbf{m}^{i+1}) \leftarrow \mathbf{v}_\eta(\mathbf{x};\mathbf{m}^i) \circ \mathbf{v}_\eta^{-1}(\mathbf{x};\delta\mathbf{m}) \tag{4.10a}$$

$$\leftarrow \mathbf{\Gamma}_\eta^{-1} \circ \mathbf{w}(\mathbf{x};\mathbf{m}^i) \circ \mathbf{w}^{-1}(\mathbf{x};\delta\mathbf{m}) \circ \mathbf{\Gamma}_\eta. \tag{4.10b}$$

In the remainder this method is called calibrated Radial Distortion KLT (cRD-KLT), which stands for calibrated KLT for RD images.

4.3.3 Extending cRD-KLT to handle Uncalibrated Images

As it will be shown in the evaluation section, the cRD-KLT is highly effective for performing image alignment of local patches in cameras with lens distortion, improving substantially the tracking accuracy and repeatability when compared with standard KLT framework. However, it has the drawback of requiring prior knowledge of the distortion parameter η, which implies a partial camera calibration.

A strategy to overcome this limitation is to use the differential image alignment to estimate both the motion and the image distortion. This passes by extending the vector $\mathbf{m}$ of model parameters in order to consider η as a unknown variable in addition to the

motion variables. In this case the warping function becomes $\mathbf{v}(\mathbf{x};\, \mathbf{q})$ with the difference with respect to $\mathbf{v}_\eta(\mathbf{x}, \mathbf{m})$ being only the vector $\mathbf{q} = (\mathbf{m}, \eta)$ of free parameters to be estimated. Unfortunately, the model $\mathbf{v}(\mathbf{x};\, \mathbf{q})$ cannot be used for image registration using inverse compositional alignment. The problem is that any vector of parameters $\mathbf{q}$ of the form $\mathbf{q} = (\mathbf{0},\, \eta)$ is a null element that turns the warping function into the identity mapping

$$\mathbf{v}(\mathbf{x};\, (\mathbf{0},\, \eta)) = \mathbf{x},\, \forall_\eta. \tag{4.11}$$

This warp does not have a simple null element because every value of η verifies the identity warp when $\mathbf{m} = \mathbf{0}$. As consequence, the Jacobian of $\mathbf{v}(\mathbf{x};\, \mathbf{q})$ evaluated for any $\mathbf{q}$ such that $\mathbf{m} = \mathbf{0}$ is singular which result in a non-invertible $\mathcal{H}_d$ that precludes the use of inverse compositional alignment. An alternative would be to use the forward additive alignment, since the only requirement needed is the differentiability of the warp with respect to the motion parameters [80, 81]. Unfortunately, the computational complexity of this approach is significantly higher than that of the efficient inverse formulation. Instead of using the forward additive alignment, the next section proposes to approximate the warp $\mathbf{v}(\mathbf{x};\, \mathbf{q})$ by assuming that the distortion is locally linear in a small neighborhood around the feature point.

4.3.4 Image Alignment in Uncalibrated Images

This section demonstrates an effective solution to avoid the singular Jacobian issue by replacing the $\mathbf{v}(\mathbf{x}; \mathbf{q})$ by a suitable approximation of the desired composed warping. As it will be experimentally shown, this approximation has minimum impact in terms of error in image registration and enables to use efficient inverse compositional alignment to estimate both motion and global image distortion in an accurate and robust manner.

Let's assume that in a small neighbourhood $\mathcal{N}$ around a feature point $\mathbf{p}$ the distortion effect can be approximated by

$$\mathbf{\Gamma}(\mathbf{x}) \approx \mathbf{\Gamma}_\mathbf{p}(\mathbf{x}; \eta) = (1 + \eta \mathbf{p}^\mathsf{T} \mathbf{p})^{-1} \mathbf{x}. \tag{4.12}$$

Remark that by replacing the radius of each point $\mathbf{x}$ by the radius of the central point

$\mathbf{p}$ of the window $\mathcal{N}$ the non-linear function $\mathbf{\Gamma}$ becomes a projective transformation $\mathbf{\Gamma}_{\mathbf{p}}(\mathbf{x};\eta)$ as shown in Fig. 4.1(b). This is a perfectly plausible approximation whenever the distance between the feature point $\mathbf{p}$ and the center of the image is substantially larger than the size of the neighborhood $\mathcal{N}$. In the situations where this is not verified, the effect of distortion is negligible, and the approximation does not introduce significant error. Replacing $\mathbf{\Gamma}$ by $\mathbf{\Gamma}_{\mathbf{p}}$ in Eq. 4.8 yields the following approximation to the ideal theoretical model (see Fig.4.1(b)):

$$\mathbf{v}_{\mathbf{p}}(\mathbf{x};\mathbf{q}) = \Big(\mathbf{\Gamma}^{-1} \circ \mathbf{w} \circ \mathbf{\Gamma}_{\mathbf{p}}\Big)(\mathbf{x};\mathbf{q}). \tag{4.13}$$

In this case, the warp has single null element, and the Jacobian is not singular when evaluated in $\mathbf{q} = \mathbf{0}$, leading to an invertible $\mathcal{H}_d$. Remark that replacing $\mathbf{\Gamma}^{-1}$ by $\mathbf{\Gamma}_{\mathbf{p}}^{-1}$ would again lead to a motion model with singular Jacobian and non-invertible $\mathcal{H}_d$. In case we aim at aligning a single image template the computation of the updated $\delta\mathbf{q}$ is similar to the cRD-KLT method. However, we have observed that typically aligning a single template ($N = 1$) provides a noisy estimation of the radial distortion for small size templates. The next section explains how to efficiently and globally estimate the distortion when $N > 1$ features are being available for registration.

Estimation of the warp parameters for $N > 1$ templates

Due to the global nature of the RD, the distortion coefficient η can be simultaneously estimated for N image templates being aligned, while keeping each the vector $\mathbf{m}$ specific for each template. Recall that we want to compute the increment $\delta\mathbf{q}$ using the inverse compositional algorithm, through the following closed-form solution:

$$\delta\mathbf{q} = \mathcal{H}_d^{-1} \sum_{\mathcal{N}} \left[\nabla \mathtt{T} \frac{\partial \mathbf{v}_{\mathbf{p}}(\mathbf{x};\mathbf{0})}{\partial \mathbf{p}}\right]^{\mathsf{T}} \Big(\mathtt{I}(\mathbf{v}_{\mathbf{p}}(\mathbf{x};\mathbf{q})) - \mathtt{T}(\mathbf{x})\Big). \tag{4.14}$$

One possible solution to globally estimate the distortion coefficient would be to perform block-by-block stacking of the feature observations and solve the system of linear equations with a standard Gaussian elimination method. Instead of doing so, we explore the sparsity of the system of linear equations for efficiently solving for the

motion parameter updates. For each image feature, Eq. 4.14 can be re-written as:

$$\mathsf{B}_{n\times n}\delta\mathbf{q}_{n\times 1} = \mathbf{e}_{n\times 1} \tag{4.15}$$

where $\mathsf{B}_{n\times n} = \mathcal{H}_d = \begin{pmatrix} \mathsf{U}_i & \mathbf{z}_i \\ \mathbf{z}_i{}^\mathsf{T} & \lambda_i \end{pmatrix}$, and n is length of the motion parameter vector $\mathbf{q}$. By performing a proper block-by-block stacking, the observation of all the N tracked features lead to the system of Eq. 4.16:

$$\begin{pmatrix} \mathsf{U}_1 & & & & \mathbf{z}_1 \\ & \mathsf{U}_2 & & & \mathbf{z}_2 \\ & & \ddots & & \vdots \\ & & & \mathsf{U}_N & \mathbf{z}_N \\ \mathbf{z}_1{}^\mathsf{T} & \mathbf{z}_2{}^\mathsf{T} & \dots & \mathbf{z}_N{}^\mathsf{T} & \lambda \end{pmatrix} \begin{pmatrix} \delta\mathbf{m}_1 \\ \delta\mathbf{m}_2 \\ \vdots \\ \delta\mathbf{m}_N \\ \delta\eta \end{pmatrix} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_N \\ d \end{pmatrix}, \tag{4.16}$$

which in short-hand can be written as

$$\begin{pmatrix} \mathsf{U} & \mathbf{z} \\ \mathbf{z}^\mathsf{T} & \lambda \end{pmatrix} \begin{pmatrix} \delta\mathbf{m} \\ \delta\eta \end{pmatrix} = \begin{pmatrix} \mathbf{e} \\ d \end{pmatrix}, \tag{4.17}$$

with $\lambda = \sum_{i=1}^{N} \lambda_i$ and $d = \sum_{i=1}^{N} d_i$. To explore the system sparsity, we perform a block-based Gaussian elimination by multiplying Eq. 4.17 on the left by $\begin{pmatrix} \mathsf{I} & 0 \\ -\mathbf{z}^\mathsf{T}\mathsf{U}^{-1} & 1 \end{pmatrix}$, which yields the following:

$$\begin{pmatrix} \mathsf{U} & \mathbf{z} \\ \mathbf{0}^\mathsf{T} & -\mathbf{z}^\mathsf{T}\mathsf{U}^{-1}\mathbf{z} + \lambda \end{pmatrix} \begin{pmatrix} \delta\mathbf{m} \\ \delta\eta \end{pmatrix} = \begin{pmatrix} \mathbf{e} \\ -\mathbf{z}^\mathsf{T}\mathsf{U}^{-1}\mathbf{e} + d \end{pmatrix} \tag{4.18}$$

where the scalar $-\mathbf{z}^\mathsf{T}\mathsf{U}^{-1}\mathbf{z}$ is the Schur complement of the matrix U [90]. The distortion parameter update $\delta\eta$ is simply computed using the following equation

$$(-\mathbf{z}^\mathsf{T}\mathsf{U}^{-1}\mathbf{z} + \lambda)\delta\eta = -\mathbf{z}^\mathsf{T}\mathsf{U}^{-1}\mathbf{e} + d \tag{4.19}$$

By taking advantage of the sparsity of the system, we end up with one more equa-

tion to solve when compared with the standard inverse alignment framework. Also relevant in terms of computational efficiency is the fact that most feature dependent blocks $(\mathsf{U}, \mathbf{z}, -\mathbf{z}^{\mathsf{T}}\mathsf{U}^{-1}\mathbf{z})$ can be computed offline. These feature-dependent blocks are recomputed only when the correspondent template is updated, with the Schur complement of U being accordingly updated. The feature-dependent motion parameters can now be estimated by simply computing the following:

$$\delta\mathbf{m}_i = \mathsf{U}_i^{-1}(\mathbf{e} - \delta\eta\,\mathbf{z}_i) \tag{4.20}$$

The inverse of U corresponds to the same computational effort of an inverse computational alignment since U_i is a 6×6 diagonal matrix that can be efficiently inverted [78, 91].

Update of the warp parameters

The final step of the algorithm concerns the update of the current parameters estimate. In theory [80, 81], the incremental warp $\mathbf{v_p}(\mathbf{x}; \delta\mathbf{q})$ must be composed with the current warp estimative. We relax this composition requirement and use an approximate relation to update the warp parameters. We start from the relation given in [80, 81]

$$\mathbf{v_p}(\mathbf{x}; \mathbf{q}^{i+1}) \leftarrow \mathbf{v_p}(\mathbf{x}; \mathbf{q}^i) \circ \mathbf{v_p}^{-1}(\mathbf{x}; \delta\mathbf{q}) \equiv \mathbf{v_p}(\mathbf{v_p}(\mathbf{x}; -\delta\mathbf{q}); \mathbf{q}^i). \tag{4.21}$$

Using this equation, we can formulate the parameters update as an additive step through the computation of a Jacobian matrix $\mathsf{J}_\mathbf{q}$ that maps the inverse compositional increment $\delta\mathbf{q}$ to its additive first-order equivalent $\mathsf{J}_\mathbf{q}\delta\mathbf{q}$ [80, 81], with the warp parameters being additively updated as $\mathbf{q}^{i+1} \leftarrow \mathbf{q}^i + \mathsf{J}_\mathbf{q}\delta\mathbf{q}$.

4.4 Calibrating Distortion with Feature Tracking

In the previous section we derived a solution for recovering the distortion in the image plane by using an image alignment framework. In feature tracking applications the template is typically small and it is important to verify if the distortion can be

decoupled from the other warp components. In this section we conduct experiment in synthetic data, where the distortion and affine motion is accurately known. Since we are working with an approximated RD compensated motion model, it is important to verify until which extend the proposed motion model remains valid for describing distortion. Moreover, it is important to verify whenever the motion components $\mathbf{m}$ tend to compensate for the distortion effect.

4.4.1 Distortion Visibility at a Low Image Level

In this experiment we study the feature influence on the distortion estimation by adding synthetic distortion to an image sequence with 20 frames. We track a variable number of features across the RD distorted sequence and compare the average RD estimation against the applied ground truth distortion.

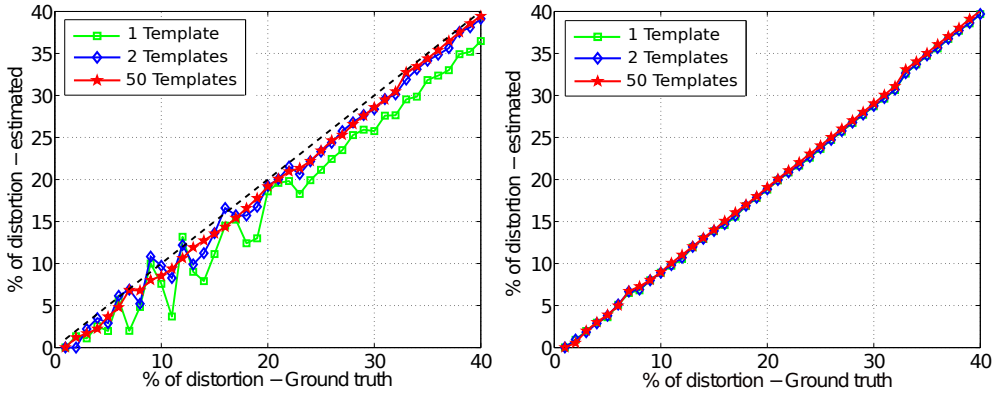
Figure 4.2 compares the distortion estimation for two integrations regions of 11 and 50 pixels. We observe that with a single feature the distortion is not very accurate because the affine motion parameters tend to compensate for the RD effect. For a moderate number of features the distortion becomes more accurate. Since the distortion is estimated globally for all the features being tracked, the affine motion models do not tend to compensate RD. It can also be observed that for large size windows, the distortion estimation is accurate, even for a single feature. It is also interesting to observe that increasing the integration regions does not largely benefit the distortion estimation for the case of 50 features.

4.4.2 Stabilizing the distortion estimation

Up to now we discuss how to estimate distortion, how to solve the sparse system of linear equations in a computational efficient manner, and the influence of the integration region and number of features in the quality of RD estimation. Now, we will show how we can integrate the estimate of the distortion parameter from each pair of images using a Kalman filter [92] to keep robust plausible distortion estimations for long-term tracking.

When deriving the equation of a Kalman filter, the goal is to find an equation that

(a) Sample sequence with RD = 30%



(b) Window Size = 11

(c) Window Size = 50

Figure 4.2: Number of features vs quality of distortion estimation. Figures show the estimation results for an integration of (b) 11 pixels and (c) and 50 pixels. For each window size we use one, two and fifty templates. The estimation using one template and small size region is very noisy, but increasing the number of templates and/or the integration region enable to accurate estimation of the distortion coefficient.

computes the *a posteriori* state estimate $\hat{\eta}_k$ as a linear combination of an *a priori* state estimate $\hat{\eta}_k^-$ and a weighted difference of an actual measurement z_k. The state of our 1-D kalman filter is the distortion coefficient η, and it is assumed to remain constant.

$$\hat{\eta}_k^- = \hat{\eta}_{k-1} \tag{4.22}$$

The measurement update equations is then the following:

$$\hat{\eta}_k = \hat{\eta}_k^- + \kappa_k(z_k - \hat{\eta}_k^-), \tag{4.23}$$

(a) 10% (b) 20% (c) 40%

Figure 4.3: Sample images of the dataset used for repeatability experiments. The images are shown with the manually selected object for controlled tracking. Inside each rectangular regular image keypoint are extract with the Shi-Tomasi detector and used for tracking evaluation purposes.

where κ_k is the kalman filter gain, z_k is the measurement that in our case is distortion estimation between two consecutive frames [92]. The Kalman estimate $\hat{\eta}_k$ is incorporated in the RD compensated warp for the next frames, with the goal now being the estimation of $\delta\eta$

$$\mathbf{v_c}(\mathbf{x};\mathbf{p},\hat{\eta}_k+\delta\eta)=\left(\mathbf{f}^{-1}\circ\mathbf{w}\circ\mathbf{g_c}\right)(\mathbf{x};\mathbf{p},\hat{\eta}_k+\delta\eta). \tag{4.24}$$

As it will be seen in the experimental validation, the application of a Kalman filter enables the stabilization of the distortion estimation in long-term tracking sequences.

4.5 Experimental Validation

We validate the alignment framework for RD images in a feature tracking context, with experiments being performed in repeatability and structure-from-motion applications. A reliable tracking algorithm must be able to perform long-term feature tracking with high pixel accuracy [75]. Typically, the tracking performance is benchmarked through the evaluation of the tracking repeatability and the spatial accuracy of the tracking [75, 82, 93]. This section compares a standard KLT algorithm against the proposed cRD-KLT and uncalibrated Radial Distortion KLT (uRD-KLT) trackers in sequences with different amounts of RD. All the trackers are directly used in the

images with distortion, without any type of rectification or pre-processing. To the best of our knowledge there is no such tracker that implicitly accounts for the effect of RD during tracking. One possible solution would be to explicitly warp the template for correcting distortion. However, this approach requires to know the RD in advance meaning that it is not a direct competitor for uRD-KLT, and introduces pernicious interpolation artifacts, so it is expected to perform worse than cRD-KLT [29, 51].

We perform experiences in sequences of planar scenes, where it is possible to obtain ground truth to assess repeatability [51, 82], and scenes with depth variation, where we evaluate the accuracy in medical endoscopy SfM [51]. The three methods under evaluation were implemented using the affine motion model and a squared integration window $\mathcal{N}$ of 11×11 inside a pyramidal image registration with $L = 4$ resolution levels. Since our main goal is to perform feature (position) tracking rather than the template itself, we monitor the health of the template through the evaluation of the squared error of Eq. 4.1, with a new template being captured at the last feature position whenever required. Note that in this case the features will not be replaced when the tracking fails since this is the exact behavior we want to evaluate.

4.5.1 Repeatability Analysis in Planar Scenes

This experiment evaluates the reliability of the feature tracking algorithms using images of planar scenes (sample images are shown in Fig. 4.3). This means that every 2 images are related by an homography that is used to verify the correctness and localization accuracy of the tracked features. For the computation of the ground truth homographies, we apply a robust estimation algorithm [49] that uses hundreds of correspondences obtained with sRD-SIFT, which provides precisely located features in radial distorted images [51]. The trackers are tested using four levels of distortion (0%, 10%, $\approx$ 20% and 40 %), with each level comprising 2 types of motion: fast translation and generic camera motion.

We start by extracting 150 features using the Shi-Tomasi detection criteria [79], and track them along the 600 frames of each sequence. The reliability of the tracks are measured using the following metrics:

Table 4.1: Performance evaluation in the planar scenes. The results are organized by type of motion (vertically) and corresponding amount of distortion (horizontally). The results presented are the RMS of the evaluation metric computed over the 600 frames.

		Fast Translation			**Affine Motion**		
		$\mathcal{R}$	$\mathcal{S}_{err}$	$\mathcal{A}_{err}$	$\mathcal{R}$	$\mathcal{S}_{err}$	$\mathcal{A}_{err}$
0%	**KLT**	0.95	0.27	0.021	0.90	0.35	0.032
	uRD-KLT	0.95	0.31	0.028	0.90	0.39	0.035
10%	**KLT**	0.92	0.58	0.055	0.90	0.59	0.045
	cRD-KLT	0.98	0.47	0.028	0.98	0.43	0.027
	uRD-KLT	0.98	0.47	0.028	0.98	0.43	0.027
20%	**KLT**	0.88	0.56	0.047	0.69	0.85	0.051
	cRD-KLT	0.98	0.43	0.026	0.90	0.55	0.027
	uRD-KLT	0.98	0.43	0.026	0.90	0.57	0.031
40%	**KLT**	0.76	1.15	0.065	0.64	1.27	0.076
	cRD-KLT	0.91	0.70	0.038	0.84	0.65	0.047
	uRD-KLT	0.90	0.73	0.040	0.84	0.65	0.047

(i) *Repeatability ($\mathcal{R}$)* measures the ratio of correct points in the frame f using the ground truth homography H_1^f that provides the mapping from view 1 to f:

$$\mathcal{R} = \frac{\#(||\mathbf{x}_f - \mathsf{H}_1^f \mathbf{x}_1||_2 < \mathcal{D})}{\#(\mathsf{H}_1^f \mathbf{x}_1)}, \tag{4.25}$$

where $||\cdot||_2$ denotes the euclidean distance and $\mathcal{D} = 2$ pixels.

(ii) *Sub-pixel accuracy ($\mathcal{S}_{err}$)* measures the RMS of the euclidean distance of correspondent feature positions as:

$$\mathcal{S}_{err} = \sqrt{\frac{\sum(||\mathbf{x}_f - \mathsf{H}_1^f \mathbf{x}_1||_2)^2}{N}}; \tag{4.26}$$

(iii) The *Photometric error* ($\mathcal{A}_{err}$) measures the RMS of the squared error of Eq. 4.1 of the N tracked features.

We also evaluate the computational time of the different methods and the RD esti-

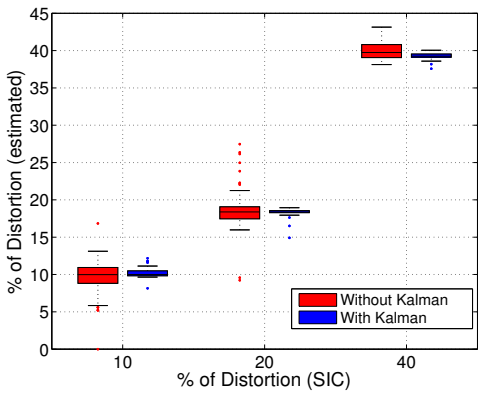


Figure 4.4: The distortion estimation is averaged over the 2 sequences with the same RD. It can be seen that using the distortion variation decreases considerably with the Kalman filter.

mation obtained using the uRD-KLT. The image sequences presenting distortion are calibrated using the single image calibration method proposed in [48, 94], which provides the ground truth for the distortion estimation.

Table 4.1 shows the repeatability results obtained in the planar image sequences. The conventional KLT tracker performs well in low distortion sequence because in this case the distortion changes smoothly between two points locations, and the template update process enables to keep plausible tracks. However, higher distortion values combined with complex motions, such as fast translation or affine camera motions, result in abrupt changes in distortion between two feature locations, precluding an effective performance of the registration process with direct consequences in the tracking results. As we increase the distortion and the complexity of the motion, the KLT starts loosing performance, which proves the importance of compensating distortion during tracking.

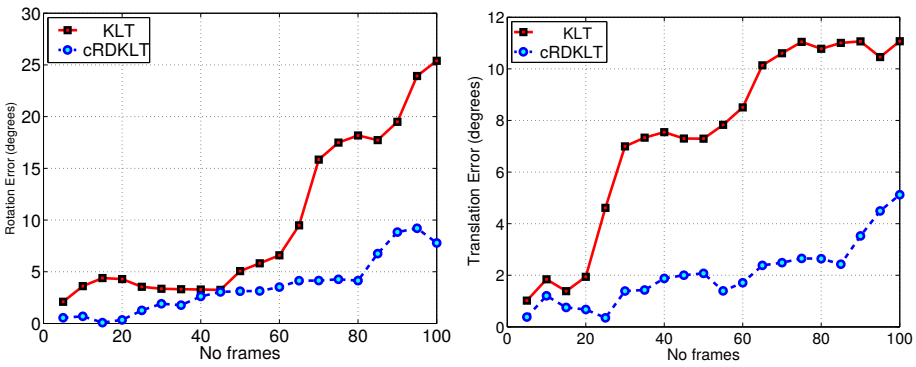
The compensation of distortion during registration, either by knowing RD calibration, or by performing it on-the-fly, brings improvements in all the evaluation parameters. The deformation tolerated by the RD compensated motion models allow to compensate the pernicious effects of distortion, which in practice is translated in accurate estimations of the feature motion parameters. This is visible in the lower appearance error and spatial accuracy achieved by the RD-KLT trackers. Since the registration is

more accurate, the appearance error is lower, and the template update is less frequent, minimizing the inherent error in localization introduced by this process. It can also be observed that uRD-KLT performs slightly worse than the cRD-KLT algorithm in the sequences with high distortion and more complex motion. The differences in sub-pixel precision and photometric error are due to the use of the approximated RD motion model, which becomes slightly more imprecise as we increase distortion. Nevertheless, the difference is almost marginal without practical influence in the repeatability. Finally, figure 4.4 shows that using the Kalman filtering reduces the variability of the distortion estimation when compared with our previous implementation in [50]

The 3 methods were implemented in Matlab/MEX files. The C-MEX files include operations that are transversal to the 3 methods, namely the interpolation routines, image gradient computation and image pyramid building. The computational times were measured in a Intel Core i7-2600 CPU @3.4GHz. cRD-KLT (≈ 1.11 milliseconds (ms)/feature) is slightly slower than the conventional KLT (≈ 1.10 ms/feature). The small differences are explained by the different motion models used, which in our case is a non-linear mapping function that requires a little more computation. The uRD-KLT (≈ 1.17 ms/feature) presents a computational overhead of $\approx 6.4\%$. Using the Schur complement instead of directly solving Eq. 4.16 enables an increasing in computational efficiency of almost 10% when compared with [50].

4.5.2 Structure-from-Motion in Medical Endoscopy

Tracking features have been successfully applied to camera motion estimation and 3D scene reconstruction [49], with accurate point correspondence across frames being of key importance for accurately recover the camera motion [9]. In the next set of experiments the motion estimation is carried by a sequential SfM pipeline that uses as input the tracked points obtained by the KLT and cRD-KLT. We have excluded the uRD-KLT from the evaluation since full camera motion calibration is required to run the adopted visual odometry pipeline. The SfM pipeline iteratively adds new consecutive frames with a 5-point RANSAC initialization (using 2 views) [52], a scale factor adjustment (using 3 views) [49], and a final refinement with a sliding window bundle adjustment.



(a) Rotation Error (b) Translation Angle error

Figure 4.5: Visual odometry evaluation. The graphics show the rotation error 4.5(a) and translation error 4.5(b). It can be seen that the stereo calibration obtained with the cRD-KLT present lower rotation and translation error, meaning that the monocular motions are more consistent than the ones obtained with the standard KLT tracker.

Visual odometry validation

The objective of this experiment is to recover the motion of a sparse sequences of 20 frames (sampled uniformly from a video sequence with 100 frames). Both trackers are initialized with the same 150 local images features, with feature replacement whenever a feature is lost. For validation purposes, we use a stereo endoscope that was calibrated $(\mathsf{R}_s\,, \mathbf{t}_s)$ with the well-known Bouguet's toolbox [1]. At each time instant we compute $(\mathsf{R}_l\,,\ \mathbf{t}_l)$ and $(\mathsf{R}_r\,,\ \mathbf{t}_r)$ by applying the visual odometry pipeline independently to the left and right channel, respectively. The computed rotations and translations are used to compute an estimative of the stereo calibration $(\mathsf{R}_s^e\,,\ \mathbf{t}_s^e)$. The rotation error is given by the angular difference between R_s^e and R_s. The translation error is evaluated by computing the angle between the two translation vectors as $\theta_t = \arccos\left(\frac{\mathbf{t}_s{}^\top \mathbf{t}_s^e}{|\mathbf{t}_s||\mathbf{t}_s^e|}\right)$.

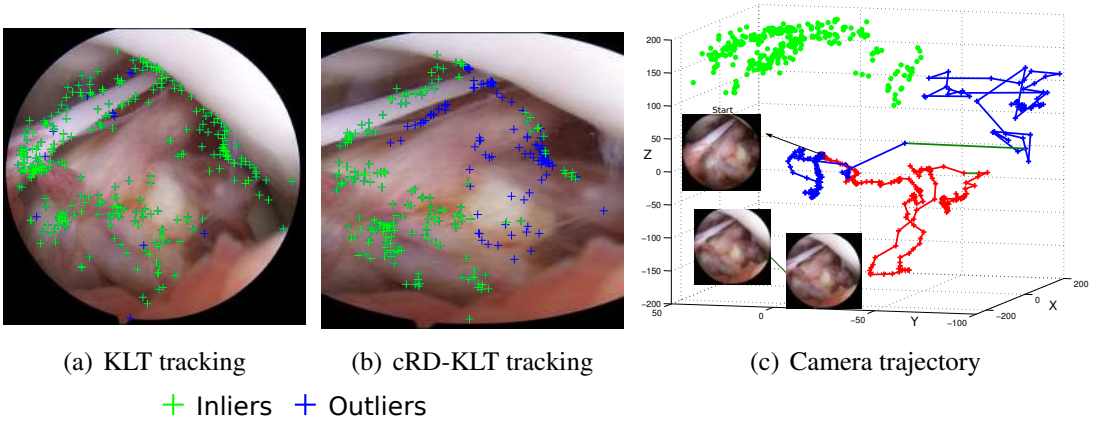
Figure 4.5 shows the evaluation of the motion estimation. By comparing the estimated stereo calibrations with the one obtained with the Bouguet toolbox, we can conclude that the cRD-KLT enable to keep consistent motion estimations on the left and right channel. The extra parameter in the RD compensated motion models permits a better convergence of the registration process in images presenting distortion,

[1] Online available at `http://www.vision.caltech.edu/bouguetj/calib_doc/`.

improving the sub-pixel accuracy of the tracked features.

In vivo validation

In this experiment we evaluate the KLT and cRD-KLT trackers in a visual odometry experiment using orthopaedic *in vivo* data acquired during a anterior cruciate ligament surgery. Since the motion is estimated between temporal adjacent images, it is expected that the camera trajectory presents smooth transitions between frames. This data set comprises 300 frames with 1920×1080 acquired at 60 fps. The high-frame rate favours the application of rigid SfM pipelines with a small bundle-adjustment window due the small deformation of the surfaces between consecutive frames. We initialize the trackers with the same 300 local images features, with feature replacement whenever a feature is lost.



(a) KLT tracking (b) cRD-KLT tracking (c) Camera trajectory

Figure 4.6: Tracking results at frame 133. Due the higher tracking precision, the cRD-KLT tracker enables to segment the non-rigid motion in the scene (classified as outlier points).Motion recover with the KLT (blue) and cRDKLT (red) trackers in the orthopaedic data set. The highlighted connection in green shows a smooth motion transition between frames 132 and 133 of the video sequence. The motion smoothness typical from continuous video is more consistent with the trajectory obtained for the cRD-KLT. The 3D structure was obtained using the cRD-KLT.

Figure 4.6 shows an example of the tracking results obtained with the KLT and cRD-KLT trackers. The final camera trajectory can be seen in Figure 4.6(c). The KLT tracked features start to drift due the combined effect of radial distortion, low-

texture and non-rigid motion, resulting in an inaccurate endoscope trajectory. Since the tracking with the cRD-KLT is more accurate, the deforming surfaces and moving tissues are more consistently removed by the visual odometry pipeline, enabling to keep a plausible trajectory estimation.

4.6 Closure

In this chapter we focus on the problem of image alignment in images presenting strong radial distortion. We improve image alignment in calibrated and uncalibrated camera setups by modifying the standard warping functions in order to account for both the motion and the non-linear image deformation arising in cameras with wide-angle lenses. Comparative experiments show that our RD-KLT tracker performs almost as well as the standard KLT tracker in sequences of correct perspective images, and achieves substantially better results in sequences with any amount of non-linear distortion. This is accomplished with minimum computational overhead. Such improvements in tracking are of strong importance for applications and domains that employ cameras equipped with mini-lens, fish-eye lenses, or boroscopes (e.g. robotics, medical applications, etc).

Chapter 5

Online Camera Zoom Calibration in Medical Endoscopy

Many image-based systems for aiding the surgeon during minimally invasive surgery require the endoscopic camera to be calibrated at all times. This chapter proposes a method for accomplishing this goal whenever the camera has optical zoom and the focal length changes during the procedure. Our solution for online calibration builds on top of the uRD-KLT for tracking salient points using differential image alignment, is well suited for continuous operation, and makes no assumptions about the camera motion or scene rigidity. Experimental validation using both a phantom model and in vivo data shows that the method enables accurate estimation of focal length when the zoom varies, avoiding the need to explicitly recalibrate during surgery. To the best of our knowledge this the first work proposing a practical solution for online zoom calibration in the OR.

5.1 Introduction

Minimally Invasive Surgery (MIS) has a number of well documented benefits for the patient, such as faster recovery time, and less trauma to surrounding tissues. However, since the surgeon has limited access to the anatomical cavity and the visualisation is carried indirectly through the video acquired by an endoscopic camera, the execution

of MIS is more difficult than the (equivalent) open-surgery. In this context, systems for Computer Assisted Surgery (CAS) that process the endoscopic video can be very helpful in assisting the doctor during the procedure, either by improving the visualisation [94], or by recovering the camera motion [19].

Most image-based CAS systems that use the endoscopic video as primary sensory input require the intrinsic camera calibration to be known at all times during the procedure [19, 94]. Endoscopic camera calibration in the context of CAS is challenging for three reasons [94, 95]: (i) since the optics are exchangeable and the camera cannot be pre-calibrated, the calibration procedure must be carried in the operation room (OR) by a non-expert user [94], (ii) in the case of oblique-viewing endoscopes the surgeon often rotates the lens scope with respect to the camera head, which changes the calibration parameters [95], and (iii) high-end endoscopy systems provide optical zoom, which means that camera focal length changes during the intervention. Melo *et al.* [94] describe effective solutions for overcoming challenges (i) and (ii). They improve usability by proposing a fully automatic calibration method that uses as input a single image of a planar checkerboard pattern and, in the case of oblique viewing endoscopes, they show that it is possible to estimate the lens rotation and update the initial calibration by tracking the image boundary contour. This chapter addresses challenge (iii) meaning that it is shown that under varying zoom the only parameter that changes significantly is the focal length, and that it is possible to update the initial calibration information without the need of re-calibrate the camera.

Zoom calibration is closely related to the problem of unknown/variable focal length estimation [96,97]. Stoyanov *et al.* [96] propose a solution for stereo endoscopy where the focal lengths are directly estimated from the fundamental matrix [37]. Assuming that the extrinsic stereo calibration is known in advance, the focal lengths can be determined using only two point matches across the stereo pair. Unfortunately, the solution only generalizes for monocular endoscopy if the camera motion is known. Stewenius *et al.* [97] propose a solution for computing the relative camera pose and unknown focal length from 6 correspondences that is used within a sample consensus framework. The method assumes a rigid scene and requires in practice a considerable baseline between images, which makes its use problematic in continuous video. Closely re-

lated to this work is the contribution of Lee *et al.* [98] that does online estimation of focal length based on the image of the boundary contour of the endoscope. Several calibrations for different zoom positions are obtained offline, and indexed using a look-up table. At running time, the boundary radius is used to index the look-up table and obtained a suitable calibration for that zoom level. This approach has the disadvantage of requiring explicit camera calibration for multiple zoom positions and, more importantly, it does not work whenever the boundary contour is not visible in the image.

5.1.1 Chapter Overview

This chapter reports a solution for efficient and accurate focal length estimation in endoscopic video. Section 5.2 presentes the adopted camera model and derives how the focal length can be estimated using global radial distortion in pixel units. Next we present a slight variation of the uRD-KLT method that enables to estimate distortion when the distortion changes between two different time instants. Since we built on tracking theory, our approach is well suited for processing continuous monocular endoscopic video, does not make assumptions about camera motion [96] or scene rigidity [97], and does not require the boundary contour of the lens to be visible [98]. Section 5.3 presents both quantitative and qualitative validation in synthetic and *in vivo* scenarios.

5.2 Zoom Calibration with the uRD-KLT

This section details the proposed method for online focal length calibration. We start by introducing the adopted camera model before moving to the method description.

5.2.1 Endoscopic Camera Modeling

Direct projection model and single image calibration

Let $\mathbf{q}_u$ be the perspective projection of a 3D point $\mathbf{Q}$ in the canonical projective plane (see Fig.5.2). In the presence of distortion, and assuming the camera to be skewless

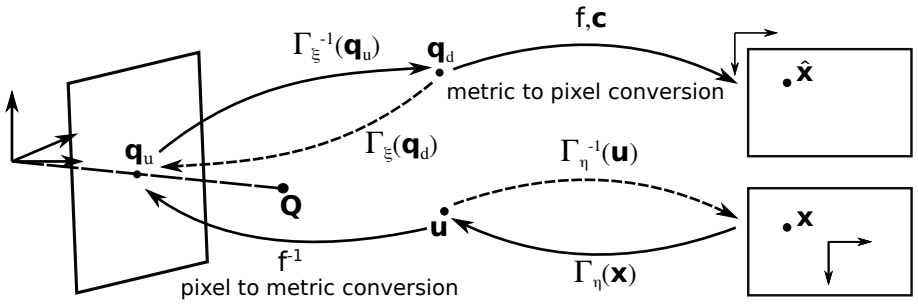


Figure 5.1: Illustration of endoscopic camera modeling in the presence of radial distortion.

and having unitary aspect ratio, point $\mathbf{q}_u$ is mapped into the point $\hat{\mathbf{x}}$ in the image plane by

$$\hat{\mathbf{x}} = f\,\mathbf{\Gamma}_\xi^{-1}(\mathbf{q}_u) \,+\, \mathbf{c}, \tag{5.1}$$

with $\mathbf{\Gamma}^{-1}(\cdot)$ being Eq. 2.6 that maps $\mathbf{q}_u$ in its distorted counterpart

$$\mathbf{q}_d = \mathbf{\Gamma}_\xi^{-1}(\mathbf{q}_u) = 2\Big(1 + \sqrt{1 - 4\xi \mathbf{q}_u{}^\top \mathbf{q}_u}\Big)^{-1} \cdot \mathbf{q}_u, \tag{5.2}$$

f is the camera focal length that converts metric units into pixel units, and $\mathbf{c} = (c_x, c_y)$ is the principal point in pixels. With the single image calibration of [94] we can easily estimate ξ, f and $\mathbf{c}$ at an initial reference zoom position. Remark that ξ is the amount of distortion in metric units that is a characteristic of the lens and therefore independent of the zoom variation.

Modeling radial distortion in the image plane

An alternative way of modelling the projection is to consider that the radial distortion acts in the image plane as opposed to act in the metric projective plane. From the inversion of Eq. 5.1 it comes in a straightforward manner that

$$\mathbf{q}_u \,=\, \mathbf{\Gamma}_\xi(f^{-1}(\hat{\mathbf{x}} \,-\, \mathbf{c})). \tag{5.3}$$

For simplicity, let's assume that $\mathbf{x} \,=\, \hat{\mathbf{x}} \,-\, \mathbf{c}$, which means that image points are expressed in a coordinate frame centred in the principal point. Replacing $\mathbf{\Gamma}_\xi$ by the

expression of Eq. 2.4 it comes that

$$f \cdot \mathbf{q}_u = (1 + \frac{\xi}{f^2}\mathbf{x}^\mathsf{T}\mathbf{x})^{-1} \cdot \mathbf{x}. \tag{5.4}$$

Let $\mathbf{u} = f \cdot \mathbf{q}_u$ be the undistorted image point in pixel units. From the equation above it follows that $\mathbf{u}$ is related with its distorted version $\mathbf{x}$ by $\mathbf{u} = \mathbf{\Gamma}_\eta(\mathbf{x})$ with

$$\eta = \xi \cdot f^{-2} \tag{5.5}$$

being the parameter that quantifies the distortion in pixel units. We conclude that, if the radial distortion is expressed in metric units, i.e. before the intrinsics, the corresponding parameter ξ does not depend of the camera focal length. However, if we quantify this same distortion in pixel units using η, then there is a dependence on the focal length which means that the distortion parameter varies with the zoom. We will use the relation of Eq. 5.5 for recovering the focal length f at each frame by combining offline calibration of the constant parameter ξ using [94] with online estimation of η using our tracking framework.

5.2.2 Zoom Calibration with Image Alignment

In the previous chapter, it is shown that it is possible to estimate the radial distortion in the image plane by tracking feature points between adjacent frames. The uRD-KLT starts by extracting reference templates $\mathtt{T}(\mathbf{x})$ around a set of salient points $\mathbf{x}$ that are detected based on image derivatives [50]. Note that in this case the previous time instant $t-1$ and the current frame at time instant t can have different distortion parameters since the distortion can change in time. Also important to note is that the distortion at instant $t-1$ is known or has been estimated. In this particular case, the deformation model is given by:

$$\mathbf{v}(\mathbf{x};\mathbf{p}) = \left(\mathbf{\Gamma}^{-1}_{\eta_{t-1}} \circ \mathbf{w} \circ \mathbf{\Gamma}\right)(\mathbf{x};\mathbf{p}), \tag{5.6}$$

with $\mathbf{p} = (\mathbf{m}, \eta_t)$ where $\mathbf{m}$ is the vector of motion parameters that describes the local deformation undergone by each image patch in the absence of distortion [81], and η is

the global distortion parameter that is common to all image regions.

Given an initial estimate of $\mathbf{p}$ the goal is to iteratively compute the updates $\delta\mathbf{p}$ of the warp parameters by minimizing the following cost function

$$\epsilon = \sum_{\mathbf{x}\in\mathcal{N}} \Big[\mathtt{I}(\mathbf{v}(\mathbf{x};\mathbf{p})) - \mathtt{T}(\mathbf{v}(\mathbf{x};\delta\mathbf{p})) \Big]^2 \tag{5.7}$$

This error function can linearised with respect to $\mathbf{p}$ by computing the first order Taylor expansion, and the final updates $\delta\mathbf{p}$ can be computed in closed-form as:

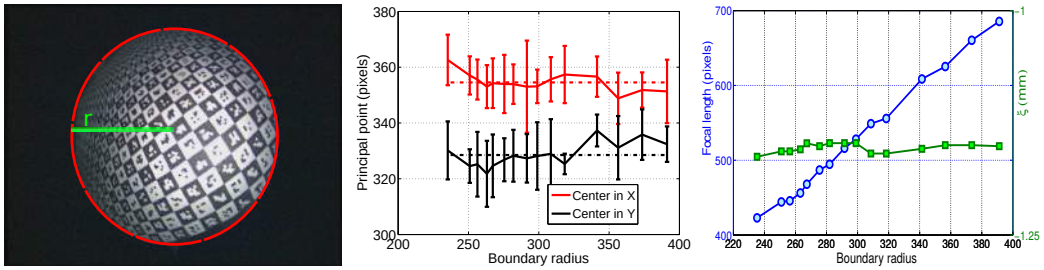
$$\delta\mathbf{p} = \mathcal{H}^{-1} \sum_{\mathbf{x}\in\mathcal{N}} \Big[\nabla\mathtt{T} \frac{\partial\mathbf{v}(\mathbf{x};\mathbf{0})}{\partial\mathbf{p}} \Big]^{\mathsf{T}} \Big(\mathtt{I}(\mathbf{v}(\mathbf{x};\mathbf{p})) - \mathtt{T}(\mathbf{x}) \Big), \tag{5.8}$$

with $\mathcal{H}$ being a 1^{st} order approximation of the Hessian matrix, and $\partial\mathbf{v}(\mathbf{x};\mathbf{0})/\partial\mathbf{p}$ being the Jacobian of the warp evaluated at the identity warp [50, 81]. Since the η is a global parameter common to every image point, the corresponding distortion updates are computed using all tracked features, while the feature local motion $\mathbf{m}$ is computed for each feature separately like detailed in the previous chapter.

While in the previous chapter, it is assumed that the camera calibration is not known and that the principal point $\mathbf{c}$ is coincident with the image center, in here we use the single image calibration [94] at a reference zoom position to obtain the principal point $\mathbf{c}$ and the lens distortion ξ in metric units. The uRD-KLT is applied during operation to continuously estimate the image distortion parameter η_t and the focal length is estimated at each frame time instant using the relation of Eq. 5.5. The approach works as far as $\mathbf{c}$ and ξ remain constant. Next section will validate the proposed method for zoom calibration, starting by empirically proving that the required assumptions hold in practice.

5.3 Experimental Validation

In this section we evaluate the proposed solution for recovering the focal length in continuous video. We start by conducting a set of experiments with ground truth to validate the assumptions made for the derivation of our solution. Afterwards, the



(a) Calibration image (b) Center of Coordinates (c) Focal length

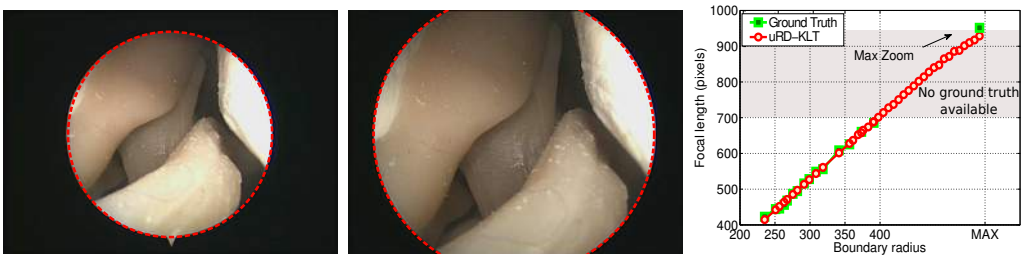
Figure 5.2: Intrinsic parameters for different zoom positions. Fig. 5.2(a) shows a calibration image where the radius of the boundary is used to index the current zoom position. Fig. 5.2(b) shows the variation of the center of coordinates and Fig. 5.2(c) shows the variation of the focal length (blue) and distortion in metric units (green) for increasing zoom. Each independent calibration is obtained using a single chessboard image. The experiment confirms that the focal length increases, while the principal point $\mathbf{c}$ and the distortion parameter ξ are virtually constant.

method is validated in both a synthetic environment and in a *in vivo* sequence acquired in a porcine uterus. To prove the usefulness of our method in real application scenarios, we include a small visual odometry experiment where the variable focal length estimation enables to correctly estimate the camera motion.

5.3.1 Variation of Intrinsic Camera Parameters with Zoom

In this experiment we used a Storz H3-Z endoscopy system with a Dyonics' arthroscopic lens with 4mm diameter. We placed the camera zoom in 15 distinct positions and, for each position, we collected 5 images of a checkerboard pattern that were used to obtain 5 independent intrinsic calibrations using the method described in [94].

Figure 5.2(b) shows the principal point estimation for successive zoom positions that are referenced using the radius of the boundary contour. Fig. 5.2(c) does the same for the focal length f and the lens distortion parameters ξ. It can be seen that all the parameters remain approximately constant with the increasing of zoom, with exception of the focal length that, as expected, increases. Therefore, the assumption that $\mathbf{c}$ and ξ are kept constant while the zoom varies holds in practice, with the variability in the center estimation being within previously reported values in the literature [48, 94].



(a) Phantom model (b) Focal length estimation

Figure 5.3: Simulation experiment with a zoom only sequence. Fig. 5.3(a) shows the two phantom images with the corresponding boundary radius at different zoom positions. Fig. 5.3(b) shows that the focal length estimation of the uRD-KLT is accurate.

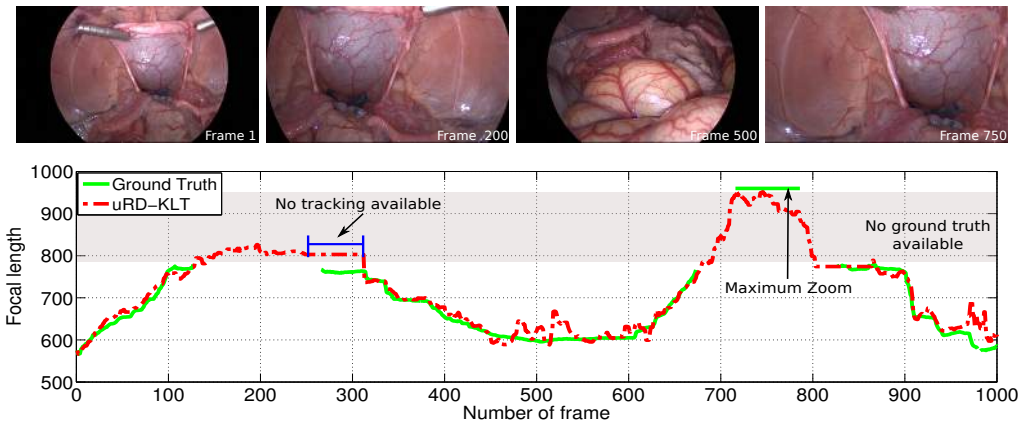
5.3.2 Validation with a Phantom Model

This experiment uses the camera setup of section 5.3.1 for acquiring a video sequence of a phantom model of the knee. The endoscope is kept stationary, while the zoom is increased. The focal length is estimated at each frame time instant by using the uRD-KLT to track 20 automatically detected points. The focal length estimates are related with the calibration results of section 5.3.1 using the radius of the boundary contour like in [98].

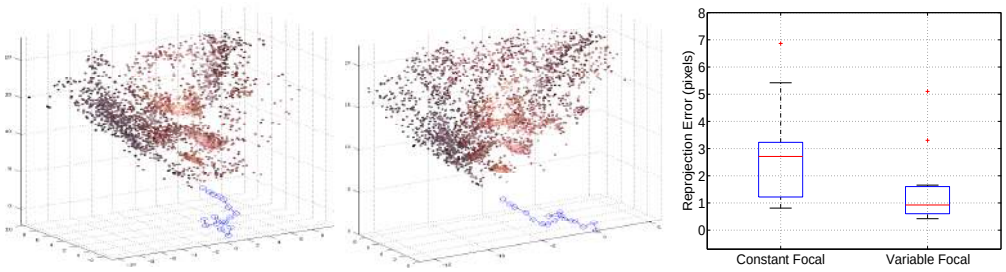
Figure 5.3 compares the on-line estimation results with the calibration ground truth. Please note that high-zoom values have no ground truth because the boundary contour is not visible and there is no manner of relating the f estimates with calibration results. Nevertheless, the estimation seems to be plausible and consistent with the calibration obtained for the end zoom position. The maximum relative estimation error was 2.5% for the maximum zoom position when the image distortion η reaches its minimum.

5.3.3 Validation in *In vivo* Data

The data used in this experiment was recorded in a *in vivo* porcine uterus during a robotic assisted procedure. The sequence of 1000 frames with resolution 1920×1080 was acquired at 30Hz with a Storz H3-Z camera system equipped with a laparoscopic lens of 10 mm from Dyonics. We used the procedure of section 5.3.1 to obtain cali-



(a) Focal length estimation



(b) Constant focal length (c) Varying focal length (d) Reprojection Error

Figure 5.4: Zoom calibration and SfM applications *in vivo data*. (a) shows some sample video frames, and the focal length estimation. (b) and (c) show a visual odometry experiment without and with focal length compensation, respectively. Compensating the focal length bring clear benefits for visual odometry, as it can be seen in Fig. 5.4(d) where the reprojection error of the reconstructed 3D points decreases from ≈ 3 pixels to less than 1 pixel.

bration ground truth. The surgeon was asked to vary the zoom against the direction of motion of the endoscope in an attempt to keep the size of the image structures constant and evaluate the robustness to changes in scale.

Figures 5.4(a) shows the online estimation results for the focal length by performing uRD-KLT tracking in the *in vivo* sequence. These results were obtained with a straightforward Matlab implementation that ran at 2 fps on a single core of an Intel i7-3630QM CPU @ 2.40GHz processor. It can be observed that the uRD-KLT-based estimation is quite accurate with an average relative error of $2.20 \pm 2.40\,\%$ when compared with the calibration ground truth. Please note that there are sequence segments

for which there are no salient points (frames 250 to 300) or the accuracy of the estimation decreases due to temporary poor tracking (frames 475-525). However, and since the focal length measurement is carried in a frame-by-frame basis, these errors do not accumulate.

Finally Fig. 5.4(b) to 5.4(d) show comparative visual odometry results for a subsequence of 17 frames where the camera moves forward while the zoom decreases. Since most of the scene is rigid the camera motion is computed by applying the five-point algorithm [52] using image correspondences obtained with sRD-SIFT [51]. The same set of points is given as input for both the case where the calibration is kept constant and varied over time, which allow to isolate the camera calibration effect on the camera motion and structure estimation. In this sequence the camera is moving towards the scene, while the zoom is removed, which from the observer point-of-view seems that the camera remains more of less stationary. This effect can be observed on the camera motion estimation with constant focal length, while for the variable focal length case the true motion pattern is captured. Figure 5.4(d) depict the motion estimation results when the focal length is kept constant and when the focal length is updated.

5.4 Closure

This chapter presents a practical solution for keeping the camera calibrated when the camera zoom changes during operation. The method builds on recent developments in image alignment for tracking keypoints in video with radial distortion and, since there are no distortion free endoscopic cameras, it can be virtually used in any MIS. The approach was validated in both synthetic and *in vivo* data, showing that is possible to keep the camera calibrated under zoom variations without the need to re-calibrate the camera during operation. To the best of our knowledge, this is the first work proposing an effective solution for the zoom calibration in continuous medical endoscopic video.

Chapter 6

Visual Odometry in Stereoscopic Laparoscopy

Stereoscopic laparoscopy provides the surgeon with depth perception at the surgical site to facilitate fine micro-manipulation of soft-tissues. The technology also enables computer-assisted laparoscopy where patient specific models can be overlaid onto laparoscopic video in real-time to provide image guidance. To maintain graphical overlay alignment of image-guides it is essential to recover the camera motion and scene geometry during the procedure. In this chapter, we propose a method for recovering the camera motion of stereo endoscopes through a multi-model fitting approach which segments rigid and non-rigid structures at the surgical site.

6.1 Introduction

Stereo laparoscopes are becoming increasingly popular in MIS. The main reason behind their wide adoption is the possibility of recovering the 3D structure of the surgical site to provide the surgeon with depth perception of the operating field. Despite of being a difficult problem due to the dynamics of the medical environment that combine occlusions from the surgical instruments with strong specularities, several authors have already proposed efficient solutions for real-time computation of depth

maps in medical endoscopy [99–102]. The obtained 3D structure can be used to align multimodal information [103] within a global reference 3D coordinate system [20] and enhance robotic instrument control.

Despite of the mature state of SfM techniques [37, 49], their application in minimally invasive surgery remains a challenging problem due to non-rigid scene deformation. An early work on structure-from-motion (SfM) in laparoscopic surgery was developed by Burschka *et al.* [19] where a rigid environment was assumed due to the confines of the sinus in order to compute a 3D scene map for registration with pre-operative Computed Tomography (CT) patient models. For procedures targeting soft-tissue anatomies non-rigidity due to cardiac, respiratory or peristaltic motions can make such SfM impossible. Deformable SfM (DSfM) [102, 104], motion compensated SLAM [54] and more recently Non-Rigid SfM [105] have been proposed for overcoming this problem but an inspection phase to build a rigid template of the scene and strong priors deformation are not always feasible. For example motion and anatomical deformation due to instrument interactions cannot be reliably modelled prior to surgery and significant practical challenges remain for robust SfM in MIS. It is also possible to incorporate position sensors [106] for additional constraints to assist the problem but this involves difficult integration solutions.

Close related work to ours was proposed by Roussos *et al.* [107] that propose a multi-body segmentation framework that uses a direct hill climbing approach to alternate the estimation of region segmentation, camera motion, and depth. This results in a computationally heavy batch algorithm that requires a quite large number of frames to become feasible. This chapter shows that by recovering depth with stereo laparoscopy the problem becomes considerably simplified, and the region segmentation and camera motion estimation can be performed online as new data arrives.

6.1.1 Chapter Overview

The next section presents a solution to effectively segment non-rigid or piecewise rigid structures from the surgical site by using multi-model fitting [22]. The method uses a temporal clustering scheme to better distinguish which scene part should be used to anchor the camera motion estimation. When compared with the state-of-the-art in

previously proposed solutions, our method does not require the entire scene to be rigid [104, 108], being robust to parts that undergo non-rigid deformation while avoiding priors on these deformations [54]. Section 6.3 presents quantitative and qualitative validation of the proposed method. Quantitative validation is performed with synthetic data [1] [99] to show the numerical stability and performance of the proposed method when the camera motion is accurately known. Qualitative validation in a long *in-vivo* video sequence shows that the proposed method is more effective in recovering the camera motion that the RANSAC-based state-of-the-art in stereo visual odometry [12].

6.2 Camera Motion Estimation in Stereo Laparoscopy

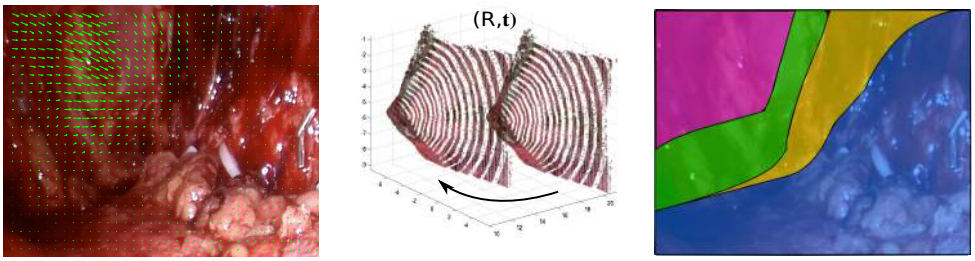
Our method can be split in three main steps: (i) computing dense correspondences between two consecutive images; (ii) generating motion hypothesis using clustering of the motion field with a multi-model fitting approach; (iii) temporal consistency based segmentation of rigid structures that enable the recovery of the camera motion. These steps are described in detail in the sections below.

6.2.1 Disparity Computation and Pixel-to-Pixel Association

The stereo endoscopic images are assumed to be rectified for disparity map computation and the device is calibrated to determine the intrinsic and extrinsic camera parameters. Given a point $\mathbf{x}_l = (x_l, y_l)^\mathsf{T}$ on the left image I_l, the goal is to compute the projection of the same point on the right image I_r that is given by $\mathbf{x}_r = (x_l + d, y_l)$. Ideally, the disparity map D is built by computing d for every image pixel. For the disparity map computation we use the method proposed by Geiger *et al.* [100] that starts by computing the disparity between control points that can be robustly matched and subsequently propagates structure into neighbouring image regions.

For associating the disparity maps between two consecutive time instants $\mathsf{D}_l \leftrightarrow \mathsf{D}'_l$ we use a standard optical flow method [110] in 2D image space $\mathsf{I}_l \leftrightarrow \mathsf{I}'_l$. For

[1] Software for rendering the synthetic data is available online at `http://www.cs.ucl.ac.uk/staff/dan.stoyanov/software.html`.



(a) Point association (b) Motion hypothesis (c) Segmented Regions

Figure 6.1: Main steps of the proposed algorithm. (a) The method starts by calculating dense correspondences in the image space. (b) Since the depths maps are available from stereo laparoscopy, the motion hypothesis are proposed in 3D using the absolute orientation method [109]. (c) The energy-based PEaRL algorithm enables to cluster the images pixels by their accordance with a certain rigid motion. Temporal consistency of the segmented regions is explored to segment rigid from non-rigid structures to solve for the camera motion.

computational reasons we do not compute the flow for every image pixel with a valid disparity and instead we sample the image space by using an equally spaced grid. Our criteria for sampling the grid is defined as function of image resolution to obtain ≈ 4000 point associations between frames.

6.2.2 Motion Hypothesis Clustering and Refinement with PEaRL

After computing the putative matches $\mathbf{x}_l \leftrightarrow \mathbf{x}'_l$, the correspondence in 3D space $\mathbf{X} \leftrightarrow \mathbf{X}'$ are obtained by using the corresponding disparity values. For registration of the 3D point clouds we use the absolute orientation method [109]. Because different motions can be present at the surgical non-rigid site, we apply the energy-based PEaRL algorithm for labelling the data points with the corresponding motion [22,111]. This procedure involves three steps: (i) generate an initial set of motion hypotheses, (ii) inlier classification by using an assigned a label (rigid motion) to the putative matches, and (iii) motion refinement using the discrete label assignment.

We start by generating camera motion hypothesis $\mathsf{T} = \begin{bmatrix} \mathsf{R} & \mathbf{t} \end{bmatrix}$ by sampling sets of 3 neighbouring points (minimal case for [109]) without repetition. Up to 500 motion hypothesis with support larger than 1% the number of pixels on the sample grid are

used. Given the set of motion hypothesis $\mathcal{T}$, the goal is to expand the models and estimate their support. This is achieved by applying PEaRL [22] to minimize the energy function

$$E(\mathbf{T}) = \underbrace{\sum_{\mathbf{x}} \mathcal{D}(\mathbf{x}, \mathsf{T}_{\mathbf{x}})}_{\text{Data cost}} + \lambda \underbrace{\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{N}} w(\mathbf{x},\mathbf{y})\delta(\mathsf{T}_{\mathbf{x}} \neq \mathsf{T}_{\mathbf{y}})}_{\text{Smoothness term}} + \underbrace{\beta|\mathcal{T}_T|}_{\text{Label cost}}, \tag{6.1}$$

where $\mathbf{T} = \{\mathsf{T}_{\mathbf{x}} | \mathbf{x} \in \mathbf{P}\}$ is an assignment of rigid motion models to data points $\mathbf{x}$.

The data cost term $\mathcal{D}(\mathbf{x}, \mathsf{T}_{\mathbf{x}})$ is the reprojection error [90] that enables to measure the error in 2D, which is more robust than directly compute the data cost in the 3D point clouds. The second term is a smoothness term that encourages the assignment of the same label (rigid motion) to spatially close point. For each data point $\mathbf{x}$ only its 10 nearest neighbours $\mathbf{y}$ are considered to compute the weight $w(\mathbf{x}, \mathbf{p})$. Since we want to enforce spatial consistency in the segmentation we consider that closer points are more likely to be described by the same rigid motion, with the weight being inversely proportional to their euclidean distance. This is achieved with the Gaussian function $w(\mathbf{x}, \mathbf{y}) = \exp\left(-||\mathbf{x} - \mathbf{y}||_2/\sigma^2\right)$. $\delta(.)$ represent the Potts model, being 1 when the condition inside parenthesis holds, and 0 otherwise [22, 111]. The label cost penalizes the number of different labels being assigned to the data points to avoid excessive fragmentation. To the possible set of rigid motions $\mathcal{T}$ we add an empty label $\emptyset$, which as a constant data cost of 1.5 pixels for all data point and label cost equal to zero. The empty label acts as outlier model, and is intended to cluster erroneous point matches, or points that are not explained by any rigid motion model hypothesis.

After the first label expansion, the motion parameters of each non-empty set are refined by using the assigned inliers. This is accomplished by minimizing the reprojection error [90] with the Levenberg-Marquardt algorithm [22, 90], with the empty labels being discarded of further optimization. The new set of labels is then used in a new expansion step with the algorithm iterating between labelling and motion refinement until the optimization does not decrease the energy of Eq. 6.1 or a certain number of iterations is reached. The constants λ and β were set to $\lambda = 1$ and $\beta = 200$. These values were empirically obtained with a independent synthetic data

set with known ground truth (not used in the method evaluation), and were used across all the experiments used along this chapter.

6.2.3 Segmenting Multi-View Consistently Labelled Parts

The minimization of the energy function of Eq. 6.1 guaranties that a label is assigned to each data point $\mathbf{x}$. Since between two consecutive frames the non-rigid or piecewise rigid structures can be subtle and easily confused with the rigid ones, we adopt a window-based system where several frames are used to perform an effective segmentation. Given a temporal window (see Fig. 6.2), we build a label-based descriptor for each pixel by concatenating the labels assigned in the frame-to-frame PEaRL optimization. Pixel descriptors with the outlier label assigned in one or more frames are discarded from further processing. The temporal segmentation is carried by clustering pixels with the exact same descriptor. In case of existing more than one cluster, the one with largest spatial support is selected as dominant rigid region and it is used to anchor the relative camera motion. Intuitively, we explore the fact that rigid structures tend to be classified with same labels in different views, the piecewise rigid or non-rigid parts tend to fragment into different labels or be classified as outliers by the PEaRL algorithm.

Finally, bundle adjustment [90] is used to refine both the camera motion and the scene structure by using only the dominant rigid part of the scene. This step is necessary because non-rigid regions can contribute on a frame-to-frame basis (locally rigid) to the optimization with PEaRL. We could apply an adaptive key frame criteria, such as the size of the segmented rigid area becoming too small, but in our current method we use a fixed four frame window for segmenting the motion to keep the running time constant. Larger temporal windows for both motion segmentation and optimization can be used at the expense of increasing the computational complexity of the algorithm.

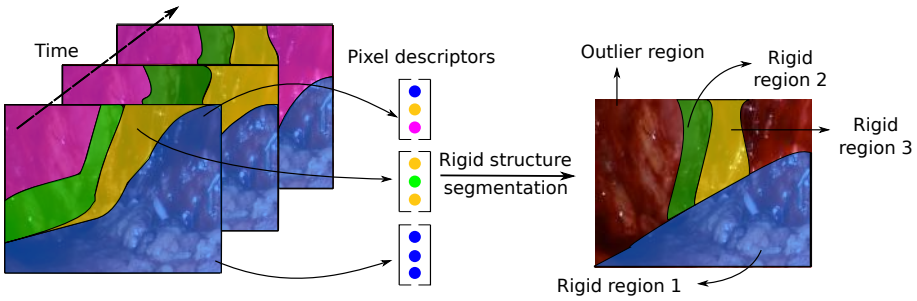


Figure 6.2: Rigid segmentation algorithm. At each frame, one label to a point correspondence (same color represent the same label, and magenta represent the outlier label). While rigid structures tend to be classified with same labels in different views, piecewise rigid or non-rigid parts tend to fragment into different labels or be classified as outliers.

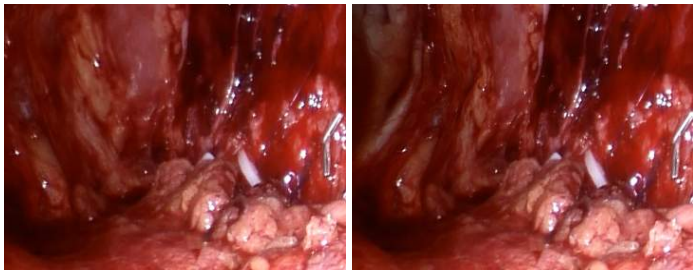
6.3 Experimental Validation

For validation of the proposed method we conduct experiments with synthetic and *in vivo* data. The proposed method was fully implemented in MATLAB, with exception of PEaRL which is implemented in C++ code [22] [2]. The single core implementation of the algorithm runs at 0.5 fps in 960×540 images on an Intel i7-3630QM CPU @ 2.40GHz processor. Our method is compared with the broadly used state-of-the-art RANSAC-based approach of [12]. This method is implemented in C++ and it runs at 2.5 fps after considerably tuning the method parameters to obtain the best possible camera motion estimations.

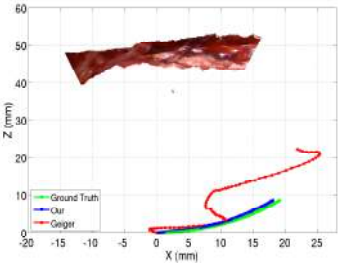
6.3.1 Experiments in Synthetic Data

Camera and scene motion ground truth is difficult to obtain for in vivo MIS video and, therefore, the proposed method is validated in a synthetic environment for which the camera motion is precisely known. While simulation sequences cannot render the full complexity of the surgical environment they allow to test the accuracy of the proposed method against different levels of white image noise to illustrate the numerical stability properties of the method. Figure 6.3 shows the performance of both methods in the simulation environment where the scene is mostly non-rigid. We

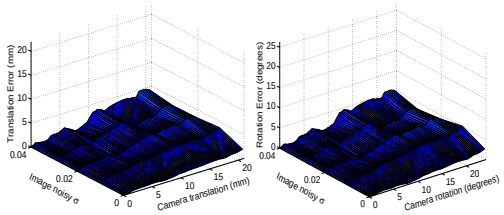
[2] Software is online available at `http://vision.csd.uwo.ca/code/`.



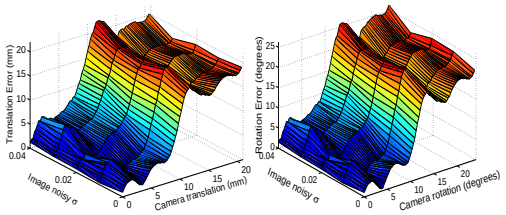
(a) Sample images



(b) Camera trajectory



(c) Our method



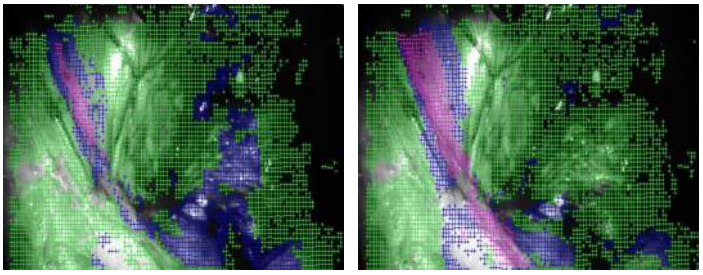
(d) Geiger's method [12]

Figure 6.3: Simulation results under increasing level of image noise. (a) show the simulation images with large deformation between them. (b) show the camera trajectory estimation for the zero noise case. Green curve represent the ground truth, blue is ours, and red is obtained with Geiger's method. (c,d) show the performance of both methods under increasing amount of additive white noise. For each method, the left graphics show the translation error as a function of the camera translation motion and level of noise. The same is done for the rotation on the right. It can be seen that our method is numerically stable under moderate levels of image noise.

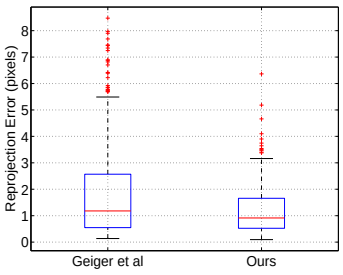
can observe that our method enables accurate camera estimation in case of such large deformation, while Geiger's method [12] tends to follow the non-rigid deformation motion.

6.3.2 Experiments in *In vivo* Data

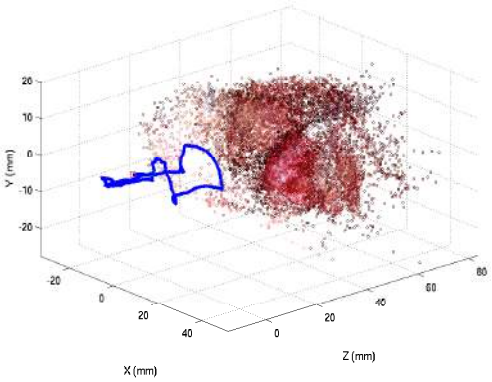
The data used in this experiment was recorded with *da Vinci Si* surgical robot during a robotically assisted prostatectomy surgery. Our and Geiger's methods [12] were used to recover the camera motion and also the dense 3D scene reconstruction. This sequence of 500 frames is particularly challenging due to the presence of non-rigid motion, strong specularities, bleeding and physiological motion due to large vascular structures in the view. At the end of the sequence the camera approximately returns to



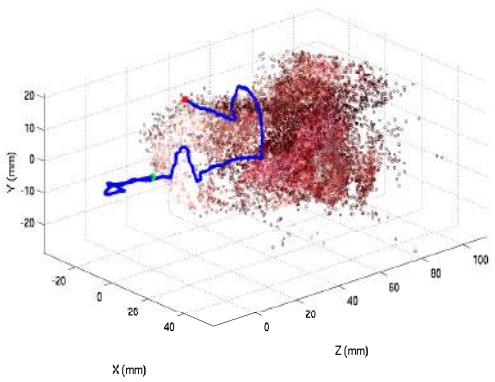
(a) Motion segmentation at frame 98 and 99



(b) Reprojection Error



(c) Our method



(d) Geiger's method [12]

Figure 6.4: Evaluation of the stereo visual odometry *in vivo* data. (a) shows two instants with the overlay segmentation. Magenta represent the outlier label that tends to increase with larger deformation. (b) shows the reprojection error obtained with each pixel in frame-by-frame basis. (c,d) show the results of our method and the method of [12] for the camera motion recovery. While our method is capable of performing reliable long-term camera motion estimation, Geiger's method [12] tends to deteriorate the estimations due to the presence the non-rigid parts. From the camera trajectories it can be easily seen that our method enables to almost close the loop.

the starting point performing a loop-closure which can be used for qualitative assessment.

Figure 6.4 shows the results for camera motion estimation using our and Geiger's methods. Since our solution effectively segments, the non-rigid parts of the scene the camera motion is reliably recovered with the rigid scene being accurately reconstructed. Geiger's method employs a conventional frame-to-frame RANSAC-based approach that is less suitable for the challenges in MIS images with the trajectory clearly drifting in the presence of non-rigid motion.

To provide a quantitative measure of the quality of the motion estimation, we compute the reprojection error in a frame-by-frame basis to show the accuracy of the camera motion estimation. While in the simulation dataset the non-rigid motion presents large amplitude and slow inter-frame variation, in this case most of the non-rigid motion is localized and very fast. The adopted segmentation criteria and window size enable good camera motion estimation in both scenarios.

6.4 Closure

This chapter presented a method for rigid structure segmentation and camera motion estimation during stereoscopic MIS. The proposed method relies on PEaRL [22] for segmenting the scene rigid structures to anchor the camera motion estimation. Temporal consistency is enforced by clustering the segmented scene structures according to the labelling assigned in the PEaRL step. Quantitative and qualitative validation in simulation and *in vivo* data show that our solution enables to keep accurate camera motion estimation in the presence of significant non-rigid deformation, outperforming the RANSAC-based state-of-the-art method in stereo visual odometry [12].

Chapter 7

Conclusions

The motivation beyond this thesis was to develop new strategies for keypoint detection, matching, and tracking in images with non-linear distortion. The proposed methods were evaluated both in medical and non-medical scenarios against state-of-the-art approaches showing that they can be applied to most vision systems that use wide FOV cameras.

In chapter 2 we have introduced the RD-SIFT and sRD-SIFT frameworks. To the best of our knowledge, we have presented the first solution for compensating the RD during feature detection and description without requiring any type of image signal resampling/interpolation. Our model-based approach implicitly introduces the radial distortion during the image scale-space computation and correct the local image gradients through a chain-rule approach. This results in a computationally efficient solution that marginally increases the computational burden when compared with the original SIFT algorithm. Repeatability experiments in non-medical scenes and SfM experiments in medical endoscopic images show that the sRD-SIFT is superior to the state-of-the-art solutions in most of the evaluation criteria. Recently, Puig *et al.* [40] evaluate the state-of-the-art methods for matching in images with RD and the sRD-SIFT has ranked first in the repeatability, matching, and run time evaluation.

In chapter 3, we show an extension of the sRD-SIFT framework to para-catadioptric images. The usefulness of this new method was demonstrated with a hybrid imaging system for indoor image-based localization. The adopted localization pipeline was

largely inspired in the object/instance retrieval literature, with descriptor vector quantization and inverted file indexing at the core of the localization engine. The final localization system outperformed by 15% a standard approach based on BOV with the standard SIFT.

Chapter 4 studies image alignment in the presence of radial distortion. We propose a generic extension to the standard motion models that describe the image template deformation in images without distortion. We study the problem of image alignment both in calibrated and uncalibrated camera setups, showing that it is possible to calibrate the distortion solely by registering local image patches. The proposed image alignment adaptations are benchmarked in feature tracking applications with repeatability experiments in non-medical scenes and SfM experiments in medical endoscopy that show their superiority against a state-of-the-art implementation of the KLT tracker.

Chapter 5 shows how to use the uRD-KLT for estimating the focal length in cameras with variable zoom. While in chapter 4, we assume the distortion is constant across all frames, in this case the distortion varies with the camera zoom, meaning that it can be different in two consecutive time instants. The pipeline results from synergies of the off-line camera calibration in [48, 94] with a slight modification to the uRD-KLT to accept variable distortion coefficients. The variability of the camera calibration parameters is studied in detail to verify that the required assumptions for the method to be feasible are verified. Evaluation in controlled experiments show that it is possible to accurately estimate the focal length by tracking features appearance across frames.

Finally, chapter 6 presents a solution for visual odometry in stereo laparoscopes. The solution is based on rigid motion segmentation and clustering using discrete optimization and temporal clustering constrains. The proposed method is compared with the state-of-the-art RANSAC approach of [12] in both simulation and *in-vivo* data showing that it is more robust in partially non-rigid scenes.

7.1 Future Work

The contributions presented in this thesis show that compensating the radial distortion during image processing can be efficiently done without requiring any type of image interpolation. Despite of the advances made there are several open issues that might worth investigate in the future.

In chapter 2 we developed a solution for feature description that enables to improve the matching performance over the state-of-the-art competitors up to 25% of distortion. One solution that can be investigated is the feature description formulation with image gradients computed in the $\mathbb{S}^2$. This provide an appropriate domain for the feature descriptor since it adapts non-linear sampling of the images and it also enhances the description step with invariance to camera relative rotation. The main challenge here would be to derive the operator on the $\mathbb{S}^2$ and devise suitable approximations for mapping the image operators into $\mathbb{P}^2$ for avoiding image interpolation.

Chapter 4 studies the problem of image alignment in radial distortion images. Although we tested the inverse composition alignment framework as a base algorithm for the registration, it might be interesting to compare it with the efficient second-order minimization proposed in [86, 87]. Mei *et al.* show that given a fixed time, the efficient second-order minimization algorithm has better convergence properties than the inverse compositional, which can be relevant for real-time applications.

Finally in chapter 6 we have proposed a framework for camera motion estimation in stereo laparoscopy in partially non-rigid environments. The algorithm starts by clustering the image points according to their rigid motion. Several frame-wise segmentations are accumulated inside a temporal window for distinguishing between the non-rigid and rigid scene structures to which the camera motion is anchored. Although in this pipeline we have selected the dominant region inside the temporal clustering window as being rigid, we think that a adaptive solution for computing the temporal window must be investigated to disambiguate situations where more than one large scene part appears to be rigid.

Bibliography

[1] C. Harris and M. Stephens, "A combined corner and edge detection," in *The Fourth Alvey Vision Conference*, 1988.

[2] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[3] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[4] H. Bay, T. Tuytelaars, and L. V. Gool, "Speeded-up Robust Features," *Computer Vision Image Understanding*, vol. 110, 2008.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *British Machine Vision Conference*, 2002.

[6] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, 2005.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization:

Improving particular object retrieval in large scale image databases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch, "Visual modeling with a hand-held camera," *International Journal of Computer Vision*, 2004.

[10] K. Wilson and N. Snavely, "Network principles for sfm: Disambiguating repeated structures with local context," in *IEEE International Conference on Computer Vision*, 2013.

[11] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *International Journal of Robotics Research*, 2002.

[12] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *Intelligent Vehicles Symposium*, 2011.

[13] W. Forstner, T. Dickscheid, and F. Schindler, "Detecting Interpretable and Accurate Scale-Invariant Keypoints," in *IEEE International Conference on Computer Vision*, 2009.

[14] P. Weinzaepfel, H. Jégou, and P. Pérez, "Reconstructing an image from its local descriptors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[15] W. Cheung and G. Hamarneh, "N-sift: N-dimensional scale invariant feature transform," *IEEE Transactions On Image Processing*, 2009.

[16] B. André, T. Vercauteren, A. M. Buchner, M. W. Shahid, M. B. Wallace, and N. Ayache, "An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis," in *Medical Image Computing and Computer-Assisted Intervention*, 2010.

[17] B. André, T. Vercauteren, A. M. Buchner, M. B. Wallace, and N. Ayache, "A Smart Atlas for Endomicroscopy using Automated Video Retrieval," *Medical Image Analysis*, 2011.

[18] T. Deselaers, H. Mller, P. Clough, H. Ney, and T. M. Lehmann, "The CLEF 2005 Automatic Medical Image Annotation Task," *International Journal of Computer Vision*, 2005.

[19] D. Burschka, M. Li, M. Ishii, R. Taylor, and G. Hager, "Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery," *Medical Image Analysis*, 2005.

[20] D. J. Mirota, H. Wang, R. H. Taylor, M. Ishii, G. L. Gallia, and G. D. Hager, "A System for Video-based Navigation for Endoscopic Endonasal Skull Base Surgery," *IEEE Transaction on Medical Imaging*, 2012.

[21] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, 1981.

[22] H. Isack and Y. Boykov, "Energy-based geometric multi-model fitting," *International Journal of Computer Vision*, 2012.

[23] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, 1999.

[24] T. Nierobisch, J. Krettek, U. Khan, and F. Hoffmann, "Optimal large view visual servoing with sets of sift features," in *IEEE International Conference on Robotics and Automation*, 2007.

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

[26] P. Baker, C. Fermuller, Y. Aloimonos, and R. Pless, "A Spherical Eye from Multiple Cameras (Makes Better Models of the World," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[27] J. Gluckman and S. Nayar, "Egomotion and Omnidirectional Cameras," in *IEEE International Conference on Computer Vision*, (Bombay), 1998.

[28] J. Barreto, J. Santos, P. Menezes, and F. Fonseca, "Ray-based Calibration of Rigid Medical Endoscopes," in *Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras*, (Marseille France), 2008.

[29] K. Daniilidis, A. Makadia, and T. Bulow, "Image Processing in Catadioptric Planes: Spaciotemporal Derivatives and Optical Flow Computation," in *International Workshop on Omndirectional Vision*, 2002.

[30] P. Hansen, P. Corke, W. Boles, and K. Daniilidis, "Scale Invariant Feature Matching with Wide Angle Images," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2007.

[31] P. Hansen, P. Corke, W. Boles, and K. Daniilidis, "Scale-Invariant Features on the Sphere," in *IEEE International Conference on Computer Vision*, 2007.

[32] P. Hansen, P. Corke, and W. Boles, "Wide-Angle Visual Feature Matching for Outdoor Localization," *International Journal of Robotics Research*, 2010.

[33] R. Castle, D. Gawley, G. Klein, and D. Murray, "Towards simultaneous recognition, localization and mapping for hand-held and wearable cameras," in *IEEE International Conference on Robotics and Automation*, 2007.

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

[35] T. Bulow, "Multiscale image processing on the sphere," in *DAGM Symposium on Pattern Recognition*, (London, UK), 2002.

[36] T. Bulow, "Spherical diffusion for 3d surface smoothing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[37] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision. 2nd edn.* Cambridge University Press, 2004.

[38] Z. Arican and P. Frossard, "OmniSIFT: Scale Invariant Features in Omnidirectional Images," in *IEEE International Conference on Image Processing*, 2010.

[39] L. Puig and J. J. Guerrero, "Scale Space for Central Catadioptric Systems. Towards a generic camera feature extractor," in *IEEE International Conference on Computer Vision*, 2011.

[40] L. Puig, K. Daniilidis, and J. Guerrero, "Scale space for camera invariant features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[41] K. Mikolajczyk, *Detection of local features invariant to affine transformations*. PhD thesis, INPG, July 2002.

[42] M. Brown and D. Lowe, "Invariant features from interest point groups," in *British Machine Vision Conference*, 2002.

[43] A. Fitzgibbon, "Simultaneous linear estimation of multiple view geometry and lens distortion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[44] J. P. Barreto, "A Unifying Geometric Representation for Central Projection Systems," *Computer Vision and Image Understading*, 2006.

[45] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, 2005.

[46] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[47] A. Haja, B. Jahne, and S. Abraham, "Localization accuracy of region detectors," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[48] J. P. Barreto, J. Roquette, P. Sturm, and F. Fonseca, "Automatic Camera Calibration Applied to Medical Endoscopy," in *British Machine Vision Conference*, 2009.

[49] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Invitation to 3D Vision: From Images to Geometric Models*. Springer-Verlag, 2003.

[50] M. Lourenco and J. Barreto, "Tracking feature points in uncalibrated images with radial distortion," in *European Conference on Computer Vision*, 2012.

[51] M. Lourenco, J. Barreto, and F. Vasconcelos, "sRD-SIFT: Keypoint Detection and Matching in Images With Radial Distortion," *IEEE Transaction Robotics*, 2012.

[52] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[53] P. Mountney, D. Stoyanov, and G.-Z. Yang, "Three-dimensional tissue deformation recovery and tracking," *IEEE Signal Processing Magazine*, 2010.

[54] P. Mountney and G.-Z. Yang, "Motion compensated slam for image guided surgery," in *Medical Image Computing and Computer-Assisted Intervention*, 2010.

[55] M. Moakher, "Means and averaging in the group of rotations," *SIAM J. Matrix Anal. Appl.*, 2002.

[56] T. Y. Tian, C. Tomasi, and D. J. Heeger, "Comparison of Approaches to Egomotion Computation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1996.

[57] S. Se, D. Lowe, and J. Little, "Vision-based Mobile Robot Localization And Mapping using Scale-Invariant Features," in *IEEE International Conference on Robotics and Automation*, 2001.

[58] A. Chavez and D. Gustafson, "Vision-Based Obstacle Avoidance Using SIFT Features," in *International Symposium on Advances in Visual Computing*, 2009.

[59] M. Cummins and P. Newman, "FAB-MAP: Probabilistic Localization and Mapping in the Space of Appearance," *International Journal of Robotics Research*, 2008.

[60] F. Werner, F. D. Maire, J. Sitte, H. Choset, S. Tully, and G. Kantor, "Topological slam using neighbourhood information of places," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2009.

[61] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[62] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003.

[63] J. K. A. C. Murillo, P. Campos and J. J. Guerrero, "GIST vocabularies in omnidirectional images for appearance based mapping and localization," in *Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras* , 2010.

[64] D. Chen, G. Baatz, K. Koeser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[65] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[66] J. Barreto and H. Araujo, "Issues on the geometry of central catadioptric image formation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[67] J. J. G. L. Puig and P. Sturm, "Matching of omnidirectional and perspective images using the hybrid fundamental matrix," in *Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras* , 2008.

[68] R. Szeliski, "Image alignment and stitching: a tutorial," *Foundations and Trends in Comput. Graph. Vis.*, 2006.

[69] J. P. Barreto and H. Araujo, "Geometric properties of central catadioptric line images and their application in calibration," *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2005.

[70] B. D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," in *DARPA Image Understanding Workshop*, 1981.

[71] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions Pattern Analysis and Machine Intelligence*, 1998.

[72] G. E. Christensen and H. J. Johnson, "Consistent image registration," *IEEE Transactions on Medical Imaging*, 2001.

[73] P. Martins, J. Batista, and R. Caseiro, "Face alignment through 2.5d active appearance models.," in *British Machine Vision Conference*, 2010.

[74] J.-Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm," 2000.

[75] M. Hwangbo, J.-S. Kim, and T. Kanade, "Gyro-aided feature tracking for a moving camera: fusion, auto-calibration and GPU implementation," *International Journal of Robotics Research*, 2011.

[76] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2011.

[77] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2011.

[78] S. J. Kim, J.-M. Frahm, and M. Pollefeys, "Joint feature tracking and radiometric calibration from auto-exposure video," in *IEEE International Conference on Computer Vision*, 2007.

[79] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

[80] S. Baker and I. Matthews, "Equivalence and Efficiency of Image Alignment Algorithms," in *IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

[81] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework'," *International Journal of Computer Vision*, 2004.

[82] S. Gauglitz, T. Höllerer, and M. Turk, "Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking," *International Journal of Computer Vision*, 2011.

[83] L. Matthews, T. Ishikawa, and S. Baker, "The Template Update Problem," *IEEE Trans. Patt. Anal. Mach. Intell.*, 2004.

[84] K. Koeser, B. Bartczak, and R. Koch, "Robust GPU-assisted camera tracking using free-form surface models," *Journal of Real-Time Image Processing*, 2007.

[85] A. Behrens, M. Bommes, T. Stehle, S. Gross, S. Leonhardt, and T. Aach, "Real-time image composition of bladder mosaics in fluorescence endoscopy," *Computer Science - Research and Development*, 2011.

[86] C. Mei, S. Benhimane, E. Malis, and P. Rives, "Efficient Homography-based Tracking and 3D Reconstruction for Single Viewpoint Sensors," *IEEE Transactions Robotics*, 2008.

[87] C. Mei, S. Benhimane, E. Malis, and P. Rives, "Constrained multiple planar template tracking for central catadioptric cameras," in *British Machine Vision Conference*, 2006.

[88] A. Salazar-Garibay, E. Malis, and C. Mei, "Visual tracking of planes with an uncalibrated central catadioptric camera," in *IROS*, 2009.

[89] T. Tamaki, T. Yamamura, and N. Ohnishi, "Unified approach to image distortion," in *IEEE International Conference on Pattern Recognition*, 2002.

[90] B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon, "Bundle adjustment a modern synthesis," in *Vision Algorithms: Theory and Practice*, 2000.

[91] G. H. Golub and C. F. van Van Loan, *Matrix Computations*. The Johns Hopkins University Press, 3rd ed., 1996.

[92] G. Welch and G. Bishop, "An introduction to the kalman filter," tech. rep., University of North Carolina at Chapel Hill, 1995.

[93] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, 2011.

[94] R. Melo, J. Barreto, and G. Falcao, "A new solution for camera calibration and real-time image distortion correction in medical endoscopy-initial technical evaluation," *IEEE Transactions on Biomedical Enginnering*, 2012.

[95] T. Yamaguchi and *et al.*, "Camera Model and Calibration Procedure for Oblique-Viewing Endoscope," in *Medical Image Computing and Computer-Assisted Intervention*, 2003.

[96] D. Stoyanov, A. Darzi, and G.-Z. Yang, "Laparoscope Self-calibration for Robotic Assisted Minimally Invasive Surgery," in *Medical Image Computing and Computer-Assisted Intervention*, 2005.

[97] H. Stewenius, D. Nister, F. Kahl, and F. Schaffalitzky, "A minimal solution for relative pose with unknown focal length," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[98] T.-Y. Lee and *et al.*, "Automatic distortion correction of endoscopic images captured with wide-angle zoom lens," *IEEE Transactions on Biomedical Engineering*, 2013.

[99] D. Stoyanov, "Stereoscopic Scene Flow for Robotic Assisted Minimally Invasive Surgery," in *Medical Image Computing and Computer-Assisted Intervention*, 2012.

[100] A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian Conference in Computer Vision*, 2010.

[101] T. Collins and A. Bartoli, "3D Reconstruction in Laparoscopy with Close-Range Photometric Stereo," in *Medical Image Computing and Computer-Assisted Intervention*, 2012.

[102] L. Maier-Hein, P. Mountney, A. Bartoli, H. Elhawary, D. Elson, A. Groch, A. Kolb, M. Rodrigues, J. Sorger, S. Speidel, and D. Stoyanov, "Optical techniques for 3d surface reconstruction in computer-assisted laparoscopic surgery," *Medical Image Analysis*, 2013.

[103] S. Roehl, S. Bodenstedt, S. Suwelack, H. Kenngott, B. P. Müller-Stich, R. Dillmann, and S. Speidel, "Dense gpu-enhanced surface reconstruction from stereo endoscopic images for intraoperative registration," *Medical Physics*, 2012.

[104] A. Malti, A. Bartoli, and T. Collins, "Template-based conformal shape-from-motion from registered laparoscopic images," in *MIUA'11*, 2011.

[105] R. Garg, A. Roussos, and L. de Agapito, "Dense variational reconstruction of non-rigid surfaces from monocular video.," in *IEEE International Conference on Computer Vision*, 2013.

[106] X. Luo and K. Mori, "Robust endoscope motion estimation via an animated particle filter for electromagnetically navigated endoscopy," *IEEE Transactions on Biomedical Engineering*, 2014.

[107] A. Roussos, C. Russell, R. Garg, and L. Agapito, "Dense multibody motion estimation and reconstruction from a handheld camera," in *IEEE International Mixed and Augmented Reality*, 2012.

[108] S. Giannarou, Z. Zhang, and G.-Z. Yang, "Deformable structure from motion by fusing visual and inertial measurement data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[109] B. K. P. Horn, "Closed-form solution of absolute orientation using unit quaternions," *Journal of the Optical Society of America*, 1987.

[110] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conference on Image Analysis*, 2003.

[111] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001.