Model-based keypoint detection in Images with Radial Distortion

Miguel Lourenço, João P. Barreto, Francisco Vasconcelos
Institute for Systems and Robotics
University of Coimbra

Abstract

Keypoint detection and matching is of fundamental importance for many applications in computer and robot vision. 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 Scale-Invariant Feature Transform (SIFT), are not resilient to the radial distortion that often arises in images acquired by cameras with micro-lenses and/or wide field-of-view. This article 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. The scale-space representation of the image is obtained using adaptive filtering that compensates the local distortion, and the keypoint description is carried after implicit image gradient correction. Unlike competing methods, our approach avoids image re-sampling (the processing is carried in the original image plane), it does not require accurate camera calibration (an approximate modeling of the distortion is sufficient), and it adds minimum computational overhead. Extensive experiments using both synthetic and real imagery show the superiority of our method in establishing point correspondence across images with radial distortion.
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 a 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 literature reports several approaches for finding image correspondences that differ in one or more of the steps enumerated above [HS88, MCUP02, Low04, BTG08]. The reader is referred to [MTS+05a, MS04, MS05, MTS+05b] for a detailed account and comparison of existing methods. The Scale-Invariant Feature Transform (SIFT), introduced by Lowe [Low04], is arguably one of the most popular matching algorithms, being broadly used in robotics for tasks like visual servoing and VSLAM [SLL02, NKKH07]. The detection is carried in a scale-space representation of the image [Lin94] that is efficiently computed using Difference-of-Gaussian (DoG). 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 robotic 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 field-of-view (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 [BFAP01, GN98]. 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 [BSMF08]. 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 can not handle the strong radial distortion introduced by the optics described above [DMB02, HCBD07a, HCBD07b, HCB10]. This article studies how radial distortion affects the performance of the original SIFT algorithm, and proposes a set of modifications that improve the detection repeatability and matching performance, while preserving the invariance to scale and rotation. The paper extends a previous conference publication that introduced RD-SIFT for the first time [LBM10]. The current journal version gives a detailed account of the method, providing a much more thorough analysis and experimental validation.

1.1 Article Overview

The radial distortion (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). Throughout the paper we will assume that the RD
can be fairly well described using the first order division model proposed in [Fit01]. Section 2 presents the camera projection model, and briefly reviews the original SIFT algorithm [Low04].

Section 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; other keypoints 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 4 and 5 suggest modifications to the SIFT framework that substantially improve its resilience to non-linear distortion. Section 4 focus on the detection, while section 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 [DMB02, HCB10], 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 5 proposes to achieve description invariance to RD by performing implicit gradient correction using the jacobian of the distortion function.

Finally, section 6 conducts several tests using real distorted images that prove the superiority and usefulness of RD-SIFT, and show that the proposed modifications preserve the original SIFT invariance to scale, rotation, illumination, and small viewpoint changes.

### 1.2 Related Work

Despite of 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 [BLTH04], others perform a preliminary correction of distortion through image rectification and then apply SIFT [CGKM07]. 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 field-of-view images should be avoided because interpolation effects introduce undesired results in filtering [DMB02]. Their article proposes using the sphere as the underlying domain of
the image function for computing optical flow in catadioptric views. However, instead of back-projecting the image plane $\mathbb{P}^2$ into the sphere $S^2$, the smoothing is formulated in $S^2$ 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 by a standard isotropic gaussian kernel that can be separated in $X$ and $Y$ [VFG08].

In [Bul02, Bul04] Bulow proposes a scale-space representation for functions defined in $S^2$ 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 [HCB07b, HCB10]. 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 $S^2$; Second, the formulation of SIFT on the sphere enables to achieve full invariance to pure camera rotation motion. The original SIFT algorithm proposed by Lowe [Low04], 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 [HZ04]. This problem can be overcome by defining the image on the sphere, for which the image features undergo rotations and translations in $S^2$.

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 $I$ into $S^2$ and convolving the result with a spherical gaussian function $G_S$ [Bul04]. 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 [DMB02, HCB07b]. We briefly review the approaches described in the literature:

- **Mapping $G_S$ into $\mathbb{P}^2$ (mSIFT):** The re-sampling can be avoided by mapping $G_S$ into
and carrying the convolution on the image plane using the original pixel values. This is similar to the adaptive filtering proposed in [DMB02], with the mapped gaussian kernel changing at every image pixel location, and the filtering not being separable in X and Y. Unfortunately this solution is unsuitable for generating the multiple scale levels of the DoG pyramid because of its computational complexity [HCBD07b,HCB10].

- **Diffusion in the Spectral Domain (sSIFT):** An alternative is to perform the gaussian smoothing in the frequency domain [HCBD07b]. Since the original image $I$ can be mapped into a spherical image $I_S$, then the spectrum of $I_S$ 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 [HCBD07b,HCB10].

- **Approximated Diffusion (pSIFT):** Hansen et al. have recently used stereographic projections for approximating the diffusion on the sphere [HCB10]. They propose to map the image $I$ 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 (separation property). 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
The RD-SIFT algorithm 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 can not 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% [HCB10]. Section 6 evaluates our algorithm in image sequences with different amounts of distortion and shows that RD-SIFT often provides twice the number of matches that would be obtained by using SIFT. 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 modeling of the distortion is sufficient).

1.3 Notation

Convolution kernels are represented by symbols in sans serif font, e.g. $G$, and image signals are denoted by symbols in typewriter font, e.g. $I$. Vectors and vector functions are typically represented by bold symbols, and scalars are indicated by plain letters, e.g $x = (x, y)^T$ and $f(x) = (f_x(x), f_y(x))^T$. We will also often use RD to refer to radial
distortion.

2 Background

2.1 The Scale Invariant Feature Transform (SIFT)

The SIFT framework was originally introduced by Lowe in [Low04]. The keypoint detection is carried in a scale-space representation of the image [Lin94], 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:

2.1.1 SIFT Detection

SIFT relies in the scale-space theory for achieving scale invariance and high repeatability in detection [Lin94]. As discussed in [Mik02], such representation requires using the Laplacian-of-Gaussian (LoG) that is the only operator that provides true scale invariance. However, Lowe [Low04] approximates the LoG by the Difference-of-Gaussian (DoG), in order to improve the computational efficiency and avoid the determination of second order derivatives that are highly sensitive to noise [VFG08]. Thus, the scale-space representation of an image $I$ is approximated by its DoG pyramid, that is computed through the subtraction of successive blurred versions of $I$. Let $L$ be a blurred version of $I$ obtained by convolution with a 2D gaussian function with standard deviation $\sigma$ (the scale).

$$L(x, y; \sigma) = I(x, y) * G(x, y; \sigma)$$ (1)
Each level of the DoG pyramid is given by

\[
\text{DoG}(x, y, k^{n+1}\sigma) = L(x, y; k^{n+1}\sigma) - L(x, y; k^n\sigma),
\]

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 \( \sigma \) [Mik02].

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 [Low04] 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 [BL02], and the position of the detected keypoints is refined to sub-pixel precision through interpolation in the DoG domain.

2.1.2 SIFT descriptor

The detection stage provides the image coordinates \( x \) and scale \( \sigma \) 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 \( \sigma \), 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 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 \( x \), and accumulates the magnitudes of the gradients whose orientations fall within its range. The gradient samples are weighted by a gaussian with center \( x \) and standard deviation 1.5\( \sigma \) 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 \( x \) and standard deviation 0.5\( \sigma \). 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 [Low04].
2.2 The Division Model for Radial Distortion

We will assume that the image distortion follows the first order division model [Fit01, Bar06], with the amount of distortion being quantified by a single parameter $\xi$ (typically $\xi < 0$), and the distortion center being approximated by the image center. Let $\mathbf{x} = (x, y)^T$ and $\mathbf{u} = (u, v)^T$ be the coordinates of corresponding points in the distorted and undistorted images expressed with respect to a reference frame with origin in the center. $\mathbf{f}$ is a vector function that maps points in the undistorted image plane $\mathcal{I}^u$ into points in the distorted image $\mathcal{I}$ [Bar06]:

$$\mathbf{x} = \mathbf{f}(\mathbf{u}) = \begin{pmatrix} f_x(u) \\ f_y(u) \end{pmatrix} = \begin{pmatrix} \frac{2u}{1 + \sqrt{1 - 4\xi(u^2 + v^2)}} \\ \frac{2v}{1 + \sqrt{1 - 4\xi(u^2 + v^2)}} \end{pmatrix}. \quad (3)$$

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

$$\mathbf{u} = \mathbf{f}^{-1}(\mathbf{x}) = \begin{pmatrix} f_u^{-1}(x) \\ f_v^{-1}(x) \end{pmatrix} = \begin{pmatrix} \frac{x}{1 + \xi(x^2 + y^2)} \\ \frac{y}{1 + \xi(x^2 + y^2)} \end{pmatrix}. \quad (4)$$

The mapping $\mathbf{f}$ 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 a particular value for the radius $r = \sqrt{x^2 + y^2}$ in the distorted image, the corresponding undistorted radius is

$$r^u = \frac{r}{1 + \xi r^2}. \quad (5)$$

The radial distortion is quantified in the division model by the parameter $\xi$. Henceforth, and in order to make the compression undergone by a particular image more intuitive, the amount of distortion will be quantified by

$$\%_{\text{distortion}} = \frac{r^u_M - r^u}{r^u} \times 100 = -\xi r M \times 100,$$

with $r_M$ denoting the distance from the center to the image corner (the maximum distorted radius).
3 Evaluation of SIFT Performance in Images with Radial Distortion

Mikolajczyk et al. 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 [Mik02, MTS+05b, MS05]. 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 equation 3. 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.

3.1 Measuring Detection Performance

Consider an image $I_0$ from the test set and its distorted version $I_d$ with $RD = d$ (see Fig. 1). Let $S_0$ and $S_d$ be respectively the sets of keypoints detected in $I_0$ and $I_d$. 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 not satisfy this invariance property. The set $S_d$ can be divided into two subsets: the set
Figure 1: The performance is evaluated on a data set comprising tenth of 640 × 480 images with different types of visual contents. The radial distortion is artificially added by warping each image using the mapping of equation 3. The figure shows one of the images of the data set to which is added increasing amounts of deformation.

\[ S_{d}^{true}, \] that comprises the keypoints that are simultaneously detected in the distorted and undistorted images

\[ S_{d}^{true} = S_{d} \cap S_{0}, \] (6)

and the set \( S_{d}^{false} \) that contains the points in \( I_{d} \) that have no correspondence in \( I_{0} \)

\[ S_{d}^{false} = S_{d} - S_{d}^{true} \]

A keypoint in the distorted image belongs to the set \( S_{d}^{true} \) iff there is a detection in \( I_{0} \) that is consistent both in space and in scale \(^{1}\). The consistency in space is verified using the mapping of equation 4. If a keypoint is detected at location \( x \) in image \( I_{d} \), then there must exist a keypoint in image \( I_{0} \) at location \( u \). In addition the scales at which the two keypoints are detected must agree. If the keypoint in the distorted image

\(^{1}\)We follow the criteria proposed in [MTS+05b, MS05] 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 [HJA08, FDS09]
has scale $\sigma_d$, then the keypoint in the undistorted image must have scale

$$\sigma_0 = \frac{\sigma_d}{1 + \xi r^2},$$

(7)

where $r$ denotes the keypoint radius in $I_d$ (see equation 5). As shown in Fig. 1, 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 $I_0$ 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

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

(8)

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(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.
3.2 Measuring Matching Performance

Assume again an image $I_0$ and one of its distorted versions $I_d$. Two keypoints are considered to be a match iff the euclidean distance between their SIFT descriptors is below a certain threshold $\lambda$ [Low04]. Let $M_d$ be the set of keypoints in $I_d$ for which the matching algorithm finds a correspondence in $I_0$. 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

\[
\text{precision} = \frac{\#M_d^{\text{true}}}{\#M_d}.
\] (9)

The precision and the recall depend on the value of the threshold \( \lambda \). In general a good matching performance is achieved whenever there is a choice for \( \lambda \) that makes both the precision and the recall close to 1. Thus, and in a similar manner to what is done in [HCBD07b, HCB10], the matching can be evaluated by verifying if the curve 1-precision Vs. recall for varying \( \lambda \) passes at a short distance of the operation point \((0, 1)\). Fig. 2(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.

### 3.3 Discussion of the Results

This section tries to interpret and understand the results observed in Fig. 2. From Fig.2(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(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(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 2.1.2). Second, the gaussian weighting of the contributions looses its efectivness. As the distortion increases, pixels in the periphery of the description region move closer to the keypoint, and contributions that would be negligible 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.

4 Improving Keypoint Detection in Images with Radial Distortion

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 due to spurious high-frequency components introduced by RD. This chapter proposes strategies for overcoming the problem. We start by discussing the benefits and drawbacks of using explicit image warping for correcting the distortion. Section
4.2 derives a new adaptive filter that compensates for the distortion while building the image scale-space representation. Section 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 4.4.

4.1 Explicit Distortion Correction using Image Warping

A straightforward 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 [CGKM07]. We employed the methodology of section 3.1, with the test frames being first distorted and then restored using image re-sampling.

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 [VFG08]. 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. 4. The left-most image is linearly re-scaled by a factor of 1.5 using bi-linear interpolation\textsuperscript{2}, 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

\textsuperscript{2}Bi-cubic interpolation was also used with similar differences in terms of detection.
Figure 3: 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 4.2 (RD-SIFT); and search for extrema in a DoG pyramid obtained using the separable approximation of the adaptive kernel derived in section 4.3 (sRD-SIFT). The evaluation is carried using the methodology of section 2 with the keypoint detection in the reference image (RD=0%) being always performed applying standard SIFT. The figures concern the repeatability 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 test sensitivity to calibration parameters were carried assuming RD = 15% detection as the ground truth. The computational time of (d) was evaluated for images (640 × 480) with increasing size and a constant distortion of 25%. This results are averaged over 15 images (all with 640 × 480 of resolution) collected on the internet, with different context, to which synthetic radial distortion was applied.
Figure 4: SIFT detection in re-sampled images. The size of the left-most image is increased in 50% using bi-linear interpolation. SIFT keypoint detection is carried independently in each frame and the results are compared. The reasons for new detections are explained in [Low04]. More surprisingly is the fact that there are keypoints in the original image that are not detected in the expanded version.

detector to find in the expanded images all the keypoints detected in the original frame. This clearly not the case.

As first argued by Daniilidis et al. [DMB02], 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. 3. 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 resampling surpass the benefits of correcting the radial distortion.

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
\begin{align}
L^u(s, t; \sigma) &= \sum_{x=-\alpha}^{\alpha} \sum_{y=-\alpha}^{\alpha} I(x, y) G\left(s - f_u^{-1}(x, y), t - f_v^{-1}(x, y); \sigma\right), \quad \text{with } \alpha = \frac{1}{\sqrt{-\xi}} \\
L(h, k; \sigma) &= \sum_{x=-\alpha}^{\alpha} \sum_{y=-\alpha}^{\alpha} I(x, y) G\left(h - x + \xi r^2(h\delta^2 - x), k - y + \xi r^2(k\delta^2 - y); \sigma\right) \tag{10}\end{align}

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 \(I\)
2. Blur the undistorted image \(I^u\) through successive convolutions with a gaussian function.
3. Apply radial distortion to the blurred images \(L^u\)
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 \(I^u\) with a gaussian kernel with standard deviation \(\sigma\). By writing the convolution operation of equation 1 explicitly, it comes that the blurred image is

\[L^u(s, t; \sigma) = \sum_{u=-\infty}^{+\infty} \sum_{v=-\infty}^{+\infty} I^u(u, v) G(s - u, t - v; \sigma).\]

If \(I\) is the original image with distortion, then it follows from section 2.2 that \(I^u(u) = I(x)\) with \(x = f(u)\) (equation 3). Replacing \(I^u\) by \(I\) and switching the variables \((u, v)\) by
\((x, y)\) using the mapping relation \(4\), we obtain the result of equation \(10\). This equation computes the undistorted blurred image \(L^u\) directly from the original distorted frame \(I\). However, it is no longer 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 \emph{adaptive convolution} that is denoted by \(\ast\) whenever convenient.

Let’s now apply radial distortion to the blurred image \(L^u\) in order to obtain \(L\). This can be achieved in an implicit manner using again the mapping relations of section 2.2. After replacing the undistorted image coordinates \((s, t)\) by their distorted counterpart \((h, k)\) and performing some algebraic simplifications, we obtain the adaptive filtering of equation \(11\) with \(r\) being the distance between the center and the image location where the filter is applied

\[
r = \sqrt{h^2 + k^2},
\]

and \(\delta\) 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 DoG pyramid that is computed by subtracting the images \(L\) of equation \(11\) for increasing values of \(\sigma\) (equation \(2\)). The new detection algorithm, henceforth dubbed \emph{RD-SIFT}, is evaluated using 15 images collected on the internet, comprising a set of different contexts. RD is synthetically added to the data set to fully isolate the RD effect during the evaluation. It can be seen in Fig. 3 that 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 almost negligible, but also such high amounts of distortion are unlikely to arise in real camera systems.
4.3 Improving Computational Efficiency

The adaptive convolution of equation 11 is computationally intensive both in terms of processing and memory requirements [DMB02]. 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 equation 11 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 \(\xi r^2\) 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 \(\delta\) is approximately unitary. Making \(\delta = 1\) in equation 11 yields

\[
\hat{L}(h, k; \sigma) = \sum_x \sum_y I(x, y) G\left(\frac{h - x}{1 + \xi r^2}, \frac{k - y}{1 + \xi r^2}; \sigma\right),
\]

with \(\hat{L}\) being an approximation of \(L\). The expression can be re-written using the adaptive convolution operator:

\[
\hat{L} = I \star \hat{G}
\]  

(13)

where \(\hat{G}\) is given by

\[
\hat{G} = G(x, y; (1 + \xi r^2)\sigma)
\]  

(14)

From the equation 14 it is easy to understand that \(I\) 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 equation 11 uses a
different kernel at every image pixel location, the approximation of equation 13 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.

It is well known that the regular 2D gaussian function $G$ can be generated by cascading two 1D gaussian kernels [Lin98, Mik02]. 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 equation 11, nor $\hat{G}$, verify the decoupling property. Despite of this let’s consider the adaptive kernel $\hat{G}$ given by

$$\hat{G} = g_h(x, y; (1 + \xi r^2)\sigma) * g_v(x, y; (1 + \xi r^2)\sigma),$$

(15)

with $g_h$ and $g_v$ being horizontal and vertical 1D gaussian functions with standard deviations varying with the radius of the convolution center. Fig. 3 evaluates the performance of the sRD-SIFT algorithm that implements the image blurring in a decoupled manner. The images $\hat{L}$, that give raise to the DoG pyramid, are obtained by the convolution of the original image $I$ with the 1D filters $g_h$ and $g_v$. 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. 3(d)).

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.
4.4.1 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. 3(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. 3(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.

4.4.2 Run Time

This experiment compares the execution time of the different detectors with respect to increasing image resolution. Fig. 3(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\textsuperscript{3}. From Fig. 3(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.

5 Improving Keypoint Description in Images with Radial Distortion

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. 5(a)). This section shows how to keep the descriptor vector stationary in order to achieve RD invariance.

\textsuperscript{3}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.
Figure 5: Matching performance by generating the SIFT descriptors before (5(a)) and after compensating for the distortion (5(b) and 5(c)). The graphics show the curves of 1-precision Vs. recall for increasing amounts of RD. In Fig. 5(b) the distortion is explicit corrected via image warping, while in Fig. 5(c) we use the implicit gradient correction approach described in 5.1.

5.1 Implicit Gradient Correction

Since we have a model for the distortion, RD invariance can be achieved by correcting the deformation before generating the descriptors. This can be done explicitly, by warping the image and computing the gradients in the undistorted signal, or implicitly, by measuring the gradients in the original image and correct the result using the derivative chain rule. The implicit approach avoids the propagation of interpolation artifacts inherent to the image re-sampling, and is computationally more efficient because the gradient correction is only performed in the description regions around the keypoints.

Let $I$ be the original image and $I^u$ be its undistorted counterpart. From section 2.2 it follows that

$$I^u(u) = I(f(u)).$$

Applying the derivative chain rule it yields

$$\nabla I^u = J_f \cdot \nabla I$$

(16)
with $\nabla I^u$ and $\nabla I$ being respectively the gradient vectors in $I^u$ and $I$, and $J_f$ being the $2 \times 2$ jacobian matrix of the mapping function $f$ given in equation 3. The Jacobian matrix can be written in terms of distorted image coordinate $x = (x,y)^T$ by replacing $u$ using the inverse mapping of equation 4:

$$J_f = \frac{1 + \xi r^2}{1 - \xi r^2} \begin{pmatrix} 1 - \xi (r^2 - 8x^2) & 8\xi xy \\ 8\xi xy & 1 - \xi (r^2 - 8y^2) \end{pmatrix}$$

with $r$ denoting the radius of $x$.

In summary, we propose to measure the gradients directly in the original distorted image $I$, evaluate the Jacobian matrix $J_f$ at every relevant pixel location, and correct the gradient vectors $\nabla I$ using equation 16. The keypoint descriptor is generated from the undistorted gradients $\nabla I^u$ following the procedure described in 2.1.1. The only modification is the replacement of the weighting gaussian function $G(x,y;\sigma)$ by the function $\hat{G} = G(x,y;(1 - \xi r^2)\sigma)$ that accounts for the changes in pixel contributions due to RD.

5.2 Evaluation in Keypoint Matching

Fig. 5 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 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 $\nabla I_u$ using equation 16 because the corresponding vector $\nabla I$ can not measured. In this case the interpolation used in the re-sampling is advantageous because it enables inferring missing information.

6 Experiments in Images with Real Lens Distortion

![Figure 6: Calibration grids and 3 (out of 13) images for each data set used for the experiments of section 6.2. The frames were acquired using a lens with low distortion (RD≈ 10%), a 4mm minilens commonly used for robotics’ applications (RD≈ 25%), and a fish-eye lens with a wide field-of-view (RD≈ 45%). The image resolution is 640 × 480 for all cases.](image)

So far all the evaluations have been carried on images with synthetic RD. This section runs comparative tests in images acquired by cameras with real lens distortion.
6.1 Methods under Evaluation

We confront the following approaches for detecting and matching keypoints between different views:

- **SIFT**: the original SIFT framework, introduced in [Low04], is directly applied to the frames with radial distortion.

- **rectSIFT**: the original SIFT is applied to the frames after correcting the distortion using image re-sampling.

- **sRD-SIFT**: a modified SIFT implementation is applied to the frames with RD. The modifications consist in using the adaptive kernel \( G \) for image blurring (section 4.3), and performing implicit gradient correction before generating the descriptors (section 5.1).

- **pSIFT**: that uses stereographic projections for approximating the diffusion on the sphere [HCB10]. The corresponding SIFT descriptor is computed by considering a support region on the sphere which is re-sampled to a canonical patch of \( 41 \times 41 \) pixels where the SIFT descriptor is computed [HCB10].

One important issue when comparing feature detectors is the variation in number of features found. Although this fact is usually ignored [MTS+05b], we try to minimize this effect by tuning the DoG threshold of the methods to provide approximately the same number of features. We tried to tune pSIFT and sRD-SIFT for the 3 image sequences. We found that it was impossible to exactly extract the same number of features and, as a complementary measure of detection performance, we have also evaluated the detectors by using the 500 strong responses over the DoG pyramid.
This first experiment uses images of planar scene surfaces acquired with the 3 lens shown in Fig. 6. All the results presented are averaged over 78 pairs of images (13 images were used in each case). The distortion center is assumed to be coincident with the image center, and the distortion parameter $\xi$ is roughly estimated by straightening up lines in the image periphery as described in [BA05]. For the pSIFT, the intrinsic calibration of each camera is obtained from a single image of a checkerboard pattern as described in [BRSF09]. Since the scenes are planar, the frames are related by an homography that can be used to verify the correctness of the matches [MTS05a, MS05]. The homographies between the images are computed in two steps. First, a small number of point correspondences are selected manually between the images. These correspondences are used to compute an homography between the images, and the second image is warped by this homography so that it is roughly aligned with the reference image. Second, a standard a robust homography estimation algorithm is used to compute an accurate residual homography between the reference and warped image (using hundreds of automatically detected and matched features). The composition of these two homographies provides the accurate ground truth homographies between the 78 possible pairs of images.

On Table 6.2 the detection step is evaluated by comparing the approaches presented. The two left-most columns show the detection time $^4$ (as defined in 4.4.2) and the total run time. Subtracting one from the other we obtain the time spent in description and matching that obviously depends on the number of detections in each image. The

\footnote{For the sRDSIFT and pSIFT all the filter were computed offline and their computation is not taken into account in this evaluation. Using the MatLab software provided by the authors [HCB10] takes in average 5 minutes to compute the octave filters required to input to pSIFT for each sequence. In Matlab implementation our filters take in average 1.25 seconds to be computed for all the octaves. In our C implementation of the method the computational overlap to the original SIFT drops to just 0.3 - 0.35 seconds.}
<table>
<thead>
<tr>
<th>TIME (sec)</th>
<th>DETECT AND MATCH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DetectTotal</td>
</tr>
<tr>
<td>10%</td>
<td>SIFT</td>
</tr>
<tr>
<td></td>
<td>rect</td>
</tr>
<tr>
<td></td>
<td>pSIFT</td>
</tr>
<tr>
<td></td>
<td>sRD</td>
</tr>
<tr>
<td>25%</td>
<td>SIFT</td>
</tr>
<tr>
<td></td>
<td>rect</td>
</tr>
<tr>
<td></td>
<td>pSIFT</td>
</tr>
<tr>
<td></td>
<td>sRD</td>
</tr>
<tr>
<td>45%</td>
<td>SIFT</td>
</tr>
<tr>
<td></td>
<td>rect</td>
</tr>
<tr>
<td></td>
<td>pSIFT</td>
</tr>
<tr>
<td></td>
<td>sRD</td>
</tr>
</tbody>
</table>

Figure 7: Detection and matching evaluation using 78 pairs of images for each level of distortion. In table 6.2 the detection statistics discussed in section 3.1. In Figs. 7(a), 7(b) and 7(c) the retrieval performances are characterized by using the sRDSIFT detector with the 4 descriptors tested. This metrics were computed by using the benchmark routines of [MS05,MTS+05b].
four columns on the right concern the keypoints detected in the first frame ($S_0$),
the common detections in both images that can be correctly associated ($S_d$), the
average repeatability ($\%Rep$) and the matching potential of the features detected by the
algorithm ($S_{true}$). sRD-SIFT and pSIFT provides by far the largest number of common
detections in both frames as well as matches. By mapping the image to the stereographic
plane via the sphere, the pSIFT adds extra rotation invariance that partially overcomes
the interpolation error of the signal resampling. The best matching potential is verified
independently of the amount of image distortion, and explains the fact that sRD-SIFT
and pSIFT always achieve a larger number of correct correspondences.

The matching performance is evaluated through the precision-recall curves exhibited
in Fig. 7(a), 7(b) and 7(c). We compare the 4 standard descriptors presented in section
6.1. For every method the precision-recall is measured over the same set of keypoints
detected using sRD-SIFT. The relative differences in results are once again consistent
with the observations of section 5.2. The implicit correction of the gradients is the best
performer for RD amounts up to 25%. Beyond 25% the explicit image correction seems
to present better resilience to the distortion effect. Surprisingly the pSIFT descriptor
present a break in terms of descriptor distinctiveness for all levels of distortion. This
fact relies on the additional re-sampling step of the support regions to a canonical patch
of $41 \times 41$ pixels for the descriptor computation [HCB10]. Although this fact might be
negligible for coarse scale features, for fine structure the interpolation interval might be
too large and inducing gross errors in the interpolated patches. This step should be
derived in the sphere and then use the mapping equations to perform the computation
in a implicit manner [DMB02].

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
Figure 8: Data set used for the SfM experiments. Each data set comprises 7 views of the scene and the results presented are averaged over the total number of combinations between views (total of 21 possible combinations).

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 [DMB02].

6.3 Structure-from-Motion (SfM)

Accurate point correspondence across frames is of key importance in multiple-view geometry [MSKS03, HZ04]. In this experiment we perform structure-from-motion using the image sequences of Fig. 8 with different amounts of RD. The intrinsic calibration
of each camera is obtained from a single image of a checkerboard pattern as described in [BRSF09]. The reference image in the first column is matched with the two other frames, and the point correspondences are used for recovering the camera motion. We apply the 5-point algorithm [Nis04] in a RANSAC procedure that estimates the epipolar geometry in a robust manner [HZ04]. Fig. 9 compares the results in recovering the camera motion and scene structure when SIFT, rectSIFT, pSIFT and sRD-SIFT are used for obtaining the image correspondences.

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 [HZ04]. We decided to vary the RANSAC threshold and compare the average number of inliers obtained with each matching algorithm. The results are shown in Fig. 9, with the RANSAC becoming more permissive and selecting more inliers as the threshold value increases. In relative terms sRD-SIFT gives raise to the largest number of inlier correspondences for every tuned threshold. This is an evidence that sRD-SIFT, not only provides more matches than competing methods, but also that it detects the keypoints in a very accurate manner (see also the experience with endoscopic images shown in the main paper). The accuracy of keypoint localization is confirmed by reprojection error obtained by our method for each image pair. For each threshold value, we perform 50 runs of RANSAC and compute the camera translation and rotation by factorizing the essential matrix [MSKS03]. The lower reprojection error reported by our method reflects the accurate sub-pixel precision of our keypoint detector.

Another important aspect is that sRD-SIFT is capable of providing high ratio of inliers (see last column of Fig. 9). This is clearly observed in this experiment since both sRD-SIFT and pSIFT provides close number of matches. However, the former method
provides a cleaner set of points (lower number of outliers). As explained in [?], this fact relies on the additional sampling of the support regions to a fixed patch size for the descriptor computation [HCB10]. Although this fact might be negligible for coarse scale features, for fine structure the interpolation interval might be too large, inducing gross errors in the interpolated patches and precluding the descriptor distinctiveness.

In summary we can conclude that sRD-SIFT gives raise to a similar number of feature matches as the pSIFT. Despite of its extra invariance to 3D camera rotation, the pSIFT methods requires several re-sampling steps via interpolation, which introcude errors both in detection and descriptor steps:

- The signal resampling to the stereographic plane introduce interpolation error that affect the sub-pixel localization that propagates to the camera motion estimate

- The pSIFT descriptor present a break in terms of descriptor distinctiveness for all levels of distortion. This fact relies on the additional re-sampling step of the support regions to a canonical patch of $41 \times 41$ pixels for the descriptor computation [HCB10]. Although this fact might be negligible for coarse scale features, for fine structure the interpolation interval might be too large and inducing gross errors in the interpolated patches

The disadvantages of pSIFT methods are compensated by the sRD-SIFT that completely avoids image interpolation.

6.4 SfM in Medical Endoscopy

This section briefly describes results in an application that was the original motivation for pursuing this work. We are currently engaged in a project that aims implementing SfM from arthroscopic images for the purpose of computer aided navigation in orthopedic surgery. The intrinsic calibration of each camera is obtained from a single image of a
checkerboard pattern as described in [BRSF09]. The reference image in the first column is matched with the others frames, and the point correspondences are used for recovering the camera motion. We apply the 5-point algorithm [Nis04] in a RANSAC [FB81] procedure that estimates the epipolar geometry in a robust manner [HZ04]. The rigid displacements between views is determined by factorization of the essential matrices [MSKS03]. In this context, finding accurate image correspondences is difficult because the endoscopic lens introduces severe radial distortion (RD ≈ 35%), and the surfaces in the joint cavity tend to be textureless (e.g. bones).

Fig. 10 compares the estimation of the rigid motion between two views of a knee joint model, when the point correspondences are obtained using SIFT, rectSIFT, pSIFT and sRD-SIFT. The scene is very poorly textured, but there are small image structures that can be potentially matched for accomplishing the task. For the case of SIFT correspondences the RANSAC procedure is unable to converge to a plausible solution for camera motion. The standard SIFT approach provides unreliable matches because it can not handle the joint effect of RD and lack of texture. The results improve when the keypoint detection and matching is carried after distortion correction via image re-sampling. However, and given the re-projection error, the accuracy of the motion estimation is not the best. The problem when using rectSIFT is that the interpolation tends to smooth the small image structures that give raise to keypoint detections. Just like before, pSIFT improves upon rectSIFT but due to signal resampling to the stereographic plane, the interpolation error affects the sub-pixel localization that propagates to the camera motion estimate. sRD-SIFT seems to handle well the situation, providing the lowest re-projection error. Since sRD-SIFT completely avoids the signal resampling it is not subject to the noise in the keypoint localization added by this step. More over the descriptor is computed over precise located keypoints providing better matching results when compared to pSIFT.
Acknowledgments

The authors acknowledge the Portuguese Science Foundation, that generously funded this work through grants PTDC/EEA-ACR/68887/2006 and SFRH/BD/63118/2009. The authors also acknowledge Peter Hansen and Peter Corke for providing the original implementation of the pSIFT algorithm. They also want to thank Abed Malti for useful discussion and support at the beginning of this project.

References


Figure 9: Structure-from-motion using point correspondences obtained with the four methods in comparison. The experiments are performed on 21 pairs of images (some images are exhibited in Fig. 8). The first column shows the average number of inliers when the RANSAC threshold increases (the average is computed after 50 independent runs of RANSAC). Second row concerns the stability in estimating the camera motion and show the average re-projection error. Since the re-projection error provides a geometric error measure of the image distance between a projected point and a measured one, it provides an excellent test to the keypoints accuracy in terms of sub-pixel localization.
Figure 10: SfM in endoscopic stereo images with low texture. The two images are overlaid and the point correspondences for each method are marked (Fig. 10(a) to 10(d)). The table shows the number of input matches for the RANSAC, the inlier selection (green matches), and the final re-projection error. The relative motion between views is refined by minimizing the re-projection error using iterative bundle adjustment [MSKS03,HZ04]. For the case of SIFT the re-projection error is not shown because the RANSAC was unable to provide a plausible initialization for the camera motion.