Noname manuscript No.
(will be inserted by the editor)

Keypoint Detection, Tracking and Matching for Medical Endoscopy

Evaluating the Impact of Radial Distortion in Medical Structure-from-Motion Applications

Miguel Lourenço · Danail Stoyanov · João P. Barreto

Received: date / Accepted: date

Abstract Minimally invasive surgeries have become standard medical procedures. In these surgeries the access to the surgical field is made by small surgical ports through which surgical instruments are inserted under the visual guidance of an endoscopic camera. Despite the number of well documented benefits for the patient, such as faster recovery time and less trauma to surrounding tissues, MIS are considerably more difficult to execute than their open-body equivalents due to the limited access to the inner body anatomies. Computer Assisted Systems (CAS) that use the endoscopic video as primary source of data have been extensively investigated to aid the surgeon during navigation. Unfortunately, most CAS invariably rely in recovering the camera motion, which in the context of medical endoscopy present additional difficulties, such as clutter and deformable surfaces, low textured surfaces, difficult illumination conditions, and strong radial distortion. This article evaluates recent methods for feature detection, matching, and tracking in the context of medical endoscopy structure-from-motion. We give special emphasis to methods that compensate for the radial distortion of medical endoscopes. These methods are evaluated using a set of benchmark evaluation criteria designed to evaluate their impact in terms of repeatability and accuracy of the camera motion estimations.

Keywords Radial Distortion · Structure-from-motion · Benchmarking · Endoscopic navigation · SIFT · KLT

1 Introduction

Minimally invasive surgeries have become standard practices, being preferable to equivalent open-body surgeries in a large set of surgeries such as ACL reconstruction. Wider dissemination of these procedures has probably been prevented by their degree of difficulty. In MIS the access to the surgical field is made by small surgical ports through which surgical instruments are inserted under visual guidance of an endoscopic camera. From the patient point-of-view, MIS procedures are highly advantageous due the faster recovery time, less trauma to the surrounding tissues and risk of post-operative complications. However, since the surgeon has limited access to the anatomical cavity and the visualization 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 CAS that process the endoscopic video can be very helpful in assisting the doctor during the procedure. The potential of image-based CAS has attracted several researchers world-wide and lead to significant advances, namely in the context of single image camera calibration [3], stereo reconstruction [8, 9, 36] and stereo visual odometry [7, 19], and instrument identification and tracking [1].

Image-based CAS invariably rely in recovering the camera motion, which in the context of medical endoscopy present several difficulties, such as clutter and

Miguel Lourenço
Vision-Box, SA, Rua do Casal de Canas, Carnaxide, Portugal
E-mail: m.lourencoeb@gmail.com

Dan Stoyanovr
Center of Medical Image Computing, University College of London, London WC1 2BT, United Kingdom
E-mail: danayl.stoyanov@ucl.ac.uk

João P. Barreto
Institute of Systems and Robotics, Faculty of Sciences and Technology, University of Coimbra
E-mail: jpbar@isr.uc.pt

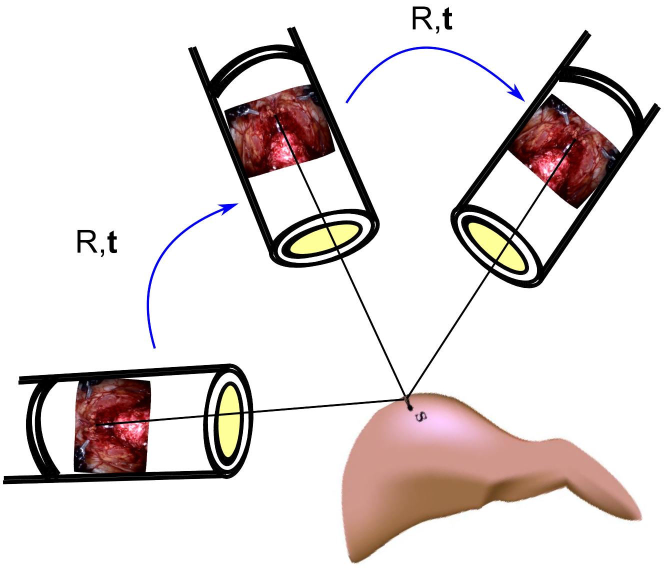


Fig. 1 Stucture-from-motion refers to the problem of computing 3D camera motion and scene structure using 2D image observations as input.

deformable surfaces, low textured surfaces, difficult illumination conditions, and strong radial distortion. Mature vision-based techniques such as structure-from-motion (SfM) and simultaneous localization and mapping have been applied in the context of endoscopic SfM applications [6, 19, 27, 38], ranging from scene registration with a pre-operative 3D model [6, 27] to robust estimation methods [38] to deal with the high percentage of outliers that typically arise when matching low texture endoscopic images. An early work on structure-from-motion (SfM) in laparoscopic surgery was developed by Burschka *et al* [6] 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 tissues, non-rigid motion due to cardiac, respiratory or peristaltic motions make such rigid SFM impossible. Deformable SfM (DSfM) [24], motion compensated SLAM [30] and more recently Non-Rigid SfM [13] have been proposed for overcoming this problem by using an inspection phase to build a rigid template of the scene and by assuming strong deformation priors . Most if not all of the SfM pipelines usually apply feature matching/tracking algorithms designed for perspective cameras [4,20,21,38] that do not account with the radial distortion (RD) effect arising in medical endoscopy imagery. Typically, this problem is overcome with a preliminary geometric correction of the radial distortion through image resampling [6, 27]. This procedure is error prune, and affects the keypoint sub-pixel accuracy due to additional image blurring introduced by interpolation techniques.

In the past few years, some authors devoted attention to the feature association problem in images acquired with cameras that do not obey the standard pinhole camera model [15, 17, 18]. Hansen *et al.* [15] proposed to map the spherical Gaussian function to an equivalent kernel on the stereographic image plane. Then, the approximate spherical diffusion is defined as the convolution of the stereographic image with the stereographic version of the Gaussian kernel. More recently, Lourenço and Barreto [18] proposed improvements to the SIFT detector/descriptors by implicitly modeling the radial distortion inside the scale-space representation and image gradient computation. More recently in [17], they developed an extension of the KLT tracker for images with strong RD, showing the advantage of using a specialized motion [12] that compensates the non-linear effect of distortion during tracking.

The main contribution of this study is the evaluation of the algorithms proposed in [15, 17, 18] in endoscopic structure-from-motion scenarios [6, 27]. Despite of the large pool of available options for feature association across views [4, 14, 20, 21, 26, 39], we use as comparative baseline the performance of the well-known SIFT [20] and KLT [21] algorithms that are still commonly used by the surgical vision community [6, 14, 27, 34, 35, 38]. We design a set of experiment that enable to quantitative evaluate the feature matching techniques with two experiments: (i) two-view geometry estimation where we evaluate sparse frame feature matching techniques [15, 18, 20], and (ii) a visual odometry experiment where we evaluate the continuous tracking algorithms of [17, 21]. As a second contribution we devise a new evaluation framework in which stereo endoscopy is used to evaluate monocular SfM pipelines, avoiding the difficult integration of optotracking devices coupled with the endoscopic camera [22, 27, 33]. Since we will be using rigid SfM pipelines for the evaluation, the non-rigid physiological motion was compensated by imaging *ex-vivo* tissues. Finally, a qualitative evaluation is performed using a visual odometry experiment using *in vivo* data collected during an orthopaedic surgery. These images present low texture and non-rigid structures.

1.1 Related Work

Image-guided MIS has received significant interest due the rapid progression in robotic assisted interventions and the evolution in imaging devices [14, 27, 34, 38, 39]. Closest references to this work are [14, 33–35, 39].

In [35] the authors propose an efficient scheme for matching keypoints detected by SIFT (or other keypoint detector) through the fitting of a local affine transformation. The algorithm starts from a set of putative matches established with SIFT descriptors according to the Euclidean distance between descriptors. Then, a hierarchical decision scheme is used for determine the affine transformation that best fits the local deformation of the surface and remove outlier matches. This work was then extended in [34] by evaluation several keypoint detectors to access which performs better for this particular task.

A probabilistic-based approach to track affine covariant regions in the context of medical endoscopy is proposed in [14]. The method starts by extract affine covariant regions and uses a EKF filter to keep plausible tracking results across time. The evaluation focus mainly on the tracking performance side and not on the accuracy of the tracking results for SfM. One key observation from this paper is that a pyramidal implementation of the KLT tracker achieves comparable performance to the proposed method [14]. Yip *et al* [39] proposed a combination of the STAR feature detector and binary robust descriptors to enable robust track features at high frame rates. This study focus on the capability of keep a tracking repeatability across frames. However, in SfM the localization precision of the keypoint is as important as the number of correct matches achieved, since keypoints with low spatial accuracy can have a negative impact in the camera motion.

To best of the our knowledge this is the first work that evaluates the reliability of feature matching techniques in recovering structure and motion in medical endoscopic scenarios. Recently, Souza *et al* [33] evaluate egomotion algorithms in the context of medical endoscopy applications. Our work focus only on the feature association step, showing that SfM accuracy can be highly improved if the non-linear deformation present on medical endoscopes is properly compensated. Studies that improve the robust estimation schemes [38], propose 2D-3D registration techniques [6, 27], or improve matching performance by means of local deformation fitting [34, 35] are complementary to ours.

1.2 Article Outline

This paper is organized as follows: Section 2 starts by reviewing the SIFT and KLT algorithm that are usually applied in medical endoscopy. This section is extended in section 2.2 by including the extensions proposed in the literature to deal with the strong radial distortion arising in unconventional optics of medical laparoscopes and endoscopes. Section 3 details the quantitative experiments conducted in this paper, with detailed information about the evaluation metrics and datasets used. Section 4 shows the experimental results, and their discussion. Finally, we draw some conclusions about the performed benchmark in section 5.

Notation: Matrices are represented by symbols in sans serif font, e.g. G, and image signals are denoted by symbols in typewriter font, e.g. $\mathtt{I}$. Vectors and vector functions are typically represented by bold symbols, and scalars are indicated by plain letters, e.g $\mathbf{x} = (x, y)^\mathsf{T}$ and $\mathbf{f}(\mathbf{x}) = (f_x(\mathbf{x}), f_y(\mathbf{x}))^\mathsf{T}$.

2 Overview of Matching Techniques

This section overviews the feature association algorithms evaluated in this paper. We start by describing the algorithms designed for perspective cameras [20, 21]. Their extensions for radial distorted images are then summarized in section 2.2

2.1 Baseline feature matching and tracking methods for perspective images

SIFT algorithm

The SIFT algorithm can be split in two different steps: keypoint detection and description. The keypoint detection uses a Difference-of-Gaussian [20] operator to perform the detection of salient points in a scale-space representation of the image [16]. Let $\mathtt{I}(\mathbf{x})$ and $\mathtt{G}(\mathbf{x};\sigma)$ be respectively an image signal and a 2D Gaussian function with standard deviation σ. The blurred version of $\mathtt{I}(x, y)$ is obtained by its convolution with the Gaussian kernel

$$\mathtt{L}(\mathbf{x};\sigma) = \mathtt{I}(\mathbf{x}) * \mathtt{G}(\mathbf{x};\sigma)\,, \tag{1}$$

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

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

Each pixel in the $\mathtt{DoG}$ pyramid is compared with its neighbours in order to find local extrema in scale and space dimensions. These extrema are subsequently filtered and refined to obtain keypoints. The next step is the computation of the descriptor vectors using the image gradients of a local patch around each detected keypoint. Scale invariance is achieved by performing all the computations at the scale of selection in the Gaussian

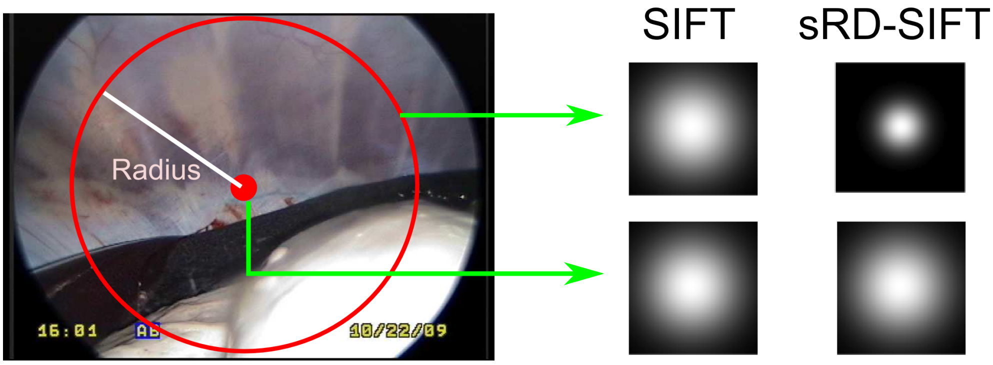


Fig. 2 Detection kernels for SIFT and sRD-SIFT. The sRD-SIFT adapts the Gaussian kernels as a functions of the image radius, enabling to adapt the deformation of the local structures caused by the radial distortion of medical endoscopes.

pyramid. The method starts by finding the dominant orientation of the local gradients, and uses it for rotating the image patch towards a normalized position. Finally, the SIFT descriptor is computed by performing a Gaussian weighting of gradient contributions, quantizing the orientations, and building histograms that accumulate magnitudes. For further details please see the original paper by Lowe [20].

Kanade-Lucas-Tomasi algorithm

Feature tracking between temporally adjacent images is typically formulated as a non-linear optimization problem whose cost function is the sum of the squared error between a template $\mathtt{T}$ and incoming images $\mathtt{I}$. The goal is to incrementally update the current motion estimate by solving:

$$\epsilon = \sum_{\mathbf{x} \in \mathcal{N}} \Big[\mathtt{I}(\mathbf{w}(\mathbf{x}; \mathbf{p} + \delta\mathbf{p})) - \mathtt{T}(\mathbf{x}) \Big]^2, \tag{3}$$

where $\mathbf{p}$ denotes the components of the image warping function $\mathbf{w}$, and $\mathcal{N}$ denotes the integration region of a feature. For efficiently solve Eq. 3, Baker and Matthews [2] proposed an *inverse compositional alignment* method that switches the roles of $\mathtt{T}$ and $\mathtt{I}$

$$\epsilon = \sum_{\mathbf{x} \in \mathcal{N}} \Big[\mathtt{I}(\mathbf{w}(\mathbf{x}; \mathbf{p})) - \mathtt{T}(\mathbf{w}(\mathbf{x}; \delta\mathbf{p})) \Big]^2. \tag{4}$$

After a Taylor expansion on $\mathbf{p}$, the motion vector increments $\delta\mathbf{p}$ can be computed as:

$$\delta\mathbf{p} = \mathcal{H}^{-1} \sum_{\mathbf{x} \in \mathcal{N}} \left[\nabla\mathtt{T} \frac{\partial \mathbf{w}(\mathbf{x}; \mathbf{0})}{\partial \mathbf{p}} \right]^{\mathsf{T}} \Big(\mathtt{I}(\mathbf{w}(\mathbf{x}; \mathbf{p})) - \mathtt{T}(\mathbf{x}) \Big), \tag{5}$$

with $\mathcal{H} = \sum_{\mathbf{x} \in \mathcal{N}} \left[\nabla\mathtt{T} \frac{\partial \mathbf{w}(\mathbf{x};\mathbf{0})}{\partial \mathbf{p}} \right]^{\mathsf{T}} \left[\nabla\mathtt{T} \frac{\partial \mathbf{w}(\mathbf{x};\mathbf{0})}{\partial \mathbf{p}} \right]$, and $\mathbf{w}(\mathbf{x}; \mathbf{0})$ being the identity warp. The computation efficiency of this inverse alignment approach relies on the dependence of $\mathcal{H}$ with the gradient templates, which means that it is constant during the registration procedure [2]. Finally, the warp parameters are updated as follows:

$$\mathbf{w}(\mathbf{x}; \mathbf{p}^{i+1}) \leftarrow \mathbf{w}(\mathbf{x}; \mathbf{p}^{i}) \circ \mathbf{w}^{-1}(\mathbf{x}; \delta\mathbf{p}). \tag{6}$$

In this paper we adopt the affine motion model [2, 17] and a pyramidal tracking framework [5], which was shown in [14] to provide good tracking results in endoscopic images.

2.2 Feature Association in Images with RD

sRD-SIFT: separable SIFT for RD images

Lourenço and Barreto proposed to use a model-based approach for image blurring that compensates for the spectral modifications caused by radial distortion [18]. The method assumes that the radial distortion can be described using the division model [11],

$$\mathbf{x} = \mathbf{f}(\mathbf{u}) = 2(1 + \sqrt{1 - 4\xi \mathbf{u}^{\mathsf{T}}\mathbf{u}}\,)^{-1}\mathbf{u}, \tag{7}$$

with $\mathbf{f}$ being a vector function that maps points from the undistorted image $\mathtt{I}^u$ to its distorted counterpart $\mathtt{I}$, $r = \sqrt{\mathbf{x}^{\mathsf{T}}\mathbf{x}}$ is the point radius with respect to the image center, and ξ is the distortion model coefficient. The scale-space image representation used in the sRD-SIFT detector algorithm [18] is built by using an approximated *distorted* Gaussian filter that enables to efficiently blur the image $\widehat{\mathtt{L}}$:

$$\widehat{\mathtt{L}}(\mathbf{h}; \sigma) = \sum_{\mathbf{x}} \mathtt{I}(\mathbf{x})\, \mathsf{G}\Big(\frac{\mathbf{h} - \mathbf{x}}{1 + \xi r^2}; \sigma \Big) \tag{8}$$

From equation 8 it follows that $\mathtt{I}$ is filtered by a Gaussian kernel with a standard deviation that varies with the image radius r. To explore the well-known separability of the Gaussian filter, the authors re-write the adaptive Gaussian kernel as being defined by:

$$\mathring{\mathtt{G}} = \mathbf{g}_h(\mathbf{x}; (1+\xi r^2)\sigma) \star \mathbf{g}_v(\mathbf{x}; (1+\xi r^2)\sigma)\,, \tag{9}$$

with $\mathbf{g}_h$ and $\mathbf{g}_v$ being horizontal and vertical 1D Gaussian functions with standard deviations varying with the radius of the convolution center. As discussed in [18], convolving the image with $\mathring{\mathtt{G}}$ is a good approximation to Eq. 8. The resulting blurred images are used to build the scale-space representation for detecting local image features, like described in section 2.1.

The image gradients for the SIFT descriptor are also corrected by using a derivative chain-rule:

$$\nabla \mathtt{I}^u = \mathsf{J}_\mathbf{f} \,.\, \nabla \mathtt{I} \tag{10}$$

with $\nabla \mathtt{I}^u$ and $\nabla \mathtt{I}$ being respectively the gradient vectors in the undistorted $\mathtt{I}^u$ and distorted $\mathtt{I}$ image signals, and $\mathsf{J}_\mathbf{f}$ being the 2×2 Jacobian matrix of the division model function $\mathbf{f}$. The process involves computing the gradients directly in the original distorted image $\mathtt{I}$, evaluate the Jacobian matrix $\mathsf{J}_\mathbf{f}$ at every relevant pixel location, and correct the gradient vectors $\nabla \mathtt{I}$ using Eq.10. For further details see [18].

pSIFT: Approximated Spherical Diffusion SIFT

Hansen et al [15] proposed to perform the Gaussian smoothing in the spectral domain. Let $\mathtt{I}_S$ be the result of back-projecting the original image I into the sphere. The spectrum of $\mathtt{I}_S$ can be found via a discrete spherical Fourier transform (DSFT), and the filtering result achieved by applying the inverse DSFT to the product of the image spectrum with the transform of G_S. Due to computational reasons, they approximate the spherical diffusion process by mapping 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 that 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. The descriptor is computed by defining a support region on the sphere and then re-sampling the region to a canonical patch of size 41×41, where the SIFT descriptor is computed. For further details see [15, 18].

cRD-KLT: calibrated KLT for images with RD

The motion models employed in KLT tracking algorithms are typically meant for perspective cameras [2]. In [17] the authors derived a composition of functions that enabled to improved the tracking in images with RD. The RD compensated motion model that relates two distorted images is expressed as follows:

$$\mathbf{x}' = \mathbf{v}_\xi(\mathbf{x};\mathbf{p}) = \Big(\mathbf{f} \circ \mathbf{w} \circ \mathbf{f}^{-1}\Big)(\mathbf{x};\mathbf{p}).,\tag{11}$$

where $\mathbf{f}$ is the division model [11] for radial distortion. Whenever the distortion is known, which is the case in this paper, the parameter vector $\mathbf{p}$ of $\mathbf{v}_\xi$ comprises the same parameters of the original motion model. Additionally, it can be proved that the requirements to be used inside the efficient inverse compositional KLT algoritm are verified [2, 17]. Therefore, the cRD-KLT consists in simply replace the proposed motion model $\mathbf{v}_\xi$ in the inverse composition KLT, being straightforward to obtain the closed form solution for $\delta\mathbf{p}$:

$$\delta\mathbf{p} = \mathcal{H}_d^{-1} \sum_{\mathbf{x}\in\mathcal{N}} \left[\nabla \mathtt{T} \frac{\partial \mathbf{v}_\xi(\mathbf{x};\mathbf{0})}{\partial \mathbf{p}}\right]^\mathsf{T} \Big(\mathtt{I}(\mathbf{v}_\xi(\mathbf{x};\mathbf{p})) - \mathtt{T}(\mathbf{x})\Big) \tag{12}$$

with $\mathcal{H}_d = \sum_{\mathbf{x}\in\mathcal{N}} \left[\nabla \mathtt{T} \frac{\partial \mathbf{v}_\xi(\mathbf{x};\mathbf{0})}{\partial \delta\mathbf{p}}\right]^\mathsf{T} \left[\nabla \mathtt{T} \frac{\partial \mathbf{v}_\xi(\mathbf{x};\mathbf{0})}{\partial \mathbf{p}}\right]$, and the Jacobian $\frac{\partial \mathbf{v}_\xi(\mathbf{x};\mathbf{0})}{\partial \mathbf{p}}$ being evaluated at $\mathbf{p} = \mathbf{0}$. Finally, the motion parameters are updated at each iteration as follows:

$$\begin{aligned} \mathbf{v}_\xi(\mathbf{x};\mathbf{p}^{i+1}) &\leftarrow \mathbf{v}_\xi(\mathbf{x};\mathbf{p}^i) \circ \mathbf{v}_\xi^{-1}(\mathbf{x};\delta\mathbf{p}) & (13)\\ &= \mathbf{f} \circ \mathbf{w}(\mathbf{x};\mathbf{p}^i) \circ \mathbf{w}^{-1}(\mathbf{x};\delta\mathbf{p}) \circ \mathbf{f}^{-1}. & (14) \end{aligned}$$

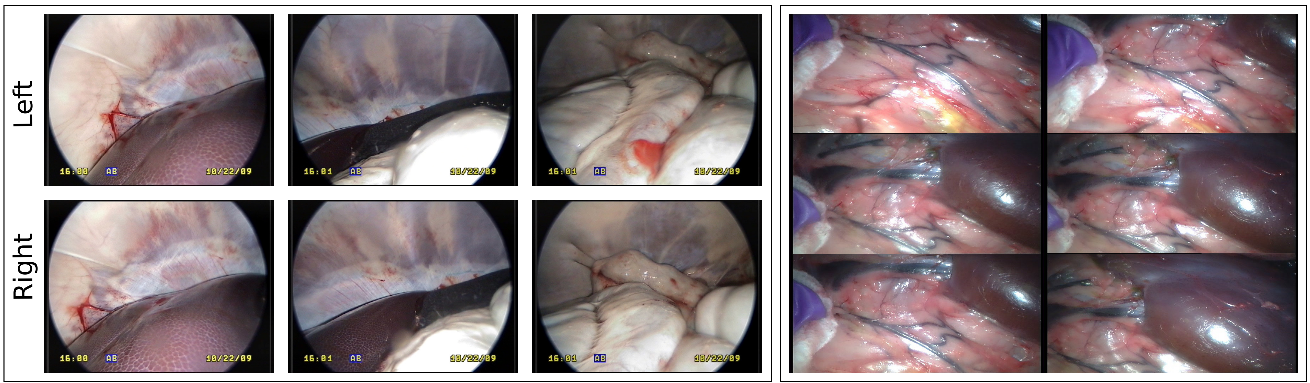
3 Evaluation Benchmarks and Endoscopic Dataset specifications

Up to now, we have summarized the theoretical details of the feature association algorithms. This section describes the benchmarks conducted along this article. We start by describing the experiment conducted, evaluation metrics, and data used for evaluating the different feature association method.

3.1 Experiment details and data

In this study we conduct two different experiments aiming at evaluating both sparse frame feature matching (SIFT, sRD-SIFT, and pSIFT) and continuous tracking algorithms.

The first experiment concerns sparse feature matching algorithms. For this experiment we consider 30 sparse image pairs of a scene with depth variation, and estimate the relative camera motion using epipolar geometry. The camera is calibrated by employing the method



(a) TVG data set

(b) VO data set

Fig. 3 Sample images used in the two experiments. Since we will be using rigid SfM pipelines for the evaluation, the non-rigid physiological motion was compensated by imaging *ex vivo* tissues

described in [3]. The rigid camera motion is estimated by the well known 5-point algorithm [32], that is run in a robust RANSAC procedure [10]. This experiment is denominated as two-view geometry (**TVG**) evaluation.

In the second experiment the objective is to recover the motion from a sequence of 100 images. The motion estimation is carried by a sequential SfM pipeline that uses as input the tracked features obtained by the KLT [2] and cRD-KLT [17] methods. The visual odometry pipeline iteratively adds new consecutive frames with a 5-point RANSAC initialization (using 2 views) [32], a scale factor adjustment (using 3 views) [23], and a final refinement with a sliding window bundle adjustment. This experiment is denominated visual odometry (**VO**) evaluation.

The datasets herein used were made available by [29,31]. The images were acquired at 25 frames per seconds (fps), with a resolution of 720 × 576 for the first experiment (see Fig. 3) and 720 × 288 for the second experiment. Due the nature of the rigid SfM experiments used in this paper, both datasets were collected by imaging *ex vivo* tissues from a porcine, which enable to minimize the non-rigid physiological motions and isolate the principal source of error we aim to evaluate, the radial distortion . The dataset used for **VO** experiment was collected with a stereo endoscope for having ground truth measures. Although, the **VO** algorithm we used is meant for monocular tracking only, we take full advantage of a pre-calibrated stereo endoscope to confirm that the two monocular camera motions are consistent. To best of our knowledge this is the first work proposing to use stereo for monocular SfM evaluation. This permits to obtain ground truth without using external optotracking devices that require additional difficult calibration procedures, like the hand-eye calibration [25, 27].

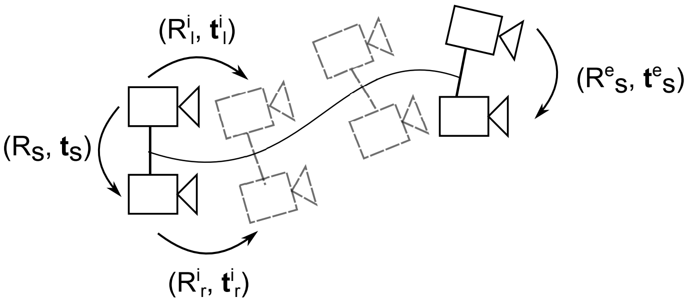

$(\mathsf{R}^i_l, \mathbf{t}^i_l)$
$(\mathsf{R}^e_S, \mathbf{t}^e_S)$
$(\mathsf{R}_S, \mathbf{t}_S)$
$(\mathsf{R}^i_r, \mathbf{t}^i_r)$

Fig. 4 Illustration of the visual odometry experiment. For validation purposes we use a stereo laparoscope, we compute the camera motion independently for each channel. The stereo calibration is used as ground truth for assessing the accuracy of the camera motion estimations.

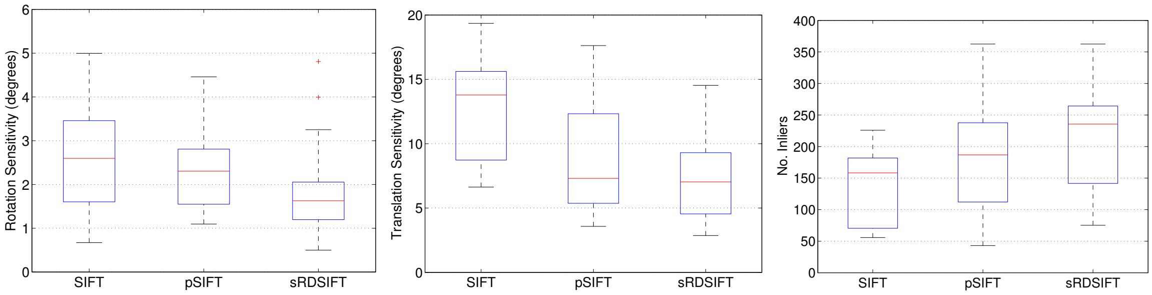
3.2 Validation metrics

For case of the **TVG** experiments we evaluate the sensitivity of motion estimates, computational time of each approach, and number of inliers retrieved by the RANSAC algorithm. The sensitivity metrics used in this experiment were the ones introduced in [37]. Given $N = 50$ trials of the RANSAC plus 5-point algorithm, we compute a *mean* rotation matrix $\bar{\mathsf{R}}$ [28] and a *mean* translation vector $\bar{\mathbf{t}}$ [37], with $\bar{\mathbf{t}}$ being a unitary vector. For each image pair, the sensitivity in translation is then computed as follows:

$$\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}[\arccos(\bar{\mathbf{t}}^{\mathsf{T}}\mathbf{t}_n)]^2}. \tag{15}$$

Like in [37], 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.

For the case of the **VO** experiment the sequence is acquired with a stereo endoscope (see Fig. refvo:example



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

Fig. 5 Two-view geometry evaluation. The graphics show the rotation 5(a) and translation 5(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.

for reference). The stereo calibration $(\mathsf{R}_s\,, \mathbf{t}_s)$ was obtained with the well-known Bouguet's toolbox[1] and it is used as ground truth. At each time instant we compute $(\mathsf{R}_l\,, \mathbf{t}_l)$ and $(\mathsf{R}_r\,, \mathbf{t}_r)$ by applying the **VO** pipeline independently to the left and right channel, respectively. At each time instant, 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 (like in the **TVG** case). The translator error is computed through the angle between the ground truth translation and estimated translated vector as $\theta_t = \arccos\left(\frac{\mathbf{t}_s{}^{\mathsf{T}}\mathbf{t}_s^e}{|\mathbf{t}_s||\mathbf{t}_s^e|}\right)$.

4 Experimental Results

In this section we present the results of the quantitative evaluation on rigid environment and qualitative evaluation *in vivo* data. Since the *in vivo* experiment is conducted using continuous video, only the KLT and cRD-KLT will be evaluated.

4.1 **TVG** experiments

Figure 5 depicts the results for the **TVG** using 30 images pairs. For the sake of visualization we combine the results of the 30 pairs using a boxplot. It can be seen in Fig. 5(c) that the sRD-SIFT algorithm enables to establish more matches than SIFT and pSIFT.

More important than 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. 5(a) and 5(b), respectively). The pSIFT algorithm improves upon SIFT in terms of number of matches obtained. However, the camera motion estimates 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 artefacts that affect the keypoint precision, which propagates to the camera motion estimation.

In terms of computational time we have observe that the SIFT algorithm outperforms the competing approaches, running at ≈ 1.21 fps. The pSIFT takes ≈ 0.67 fps, without taking into account the stereographic filters computation, which we have performed offline and took approximately ≈ 10 seconds with the MATLAB implementation provided by the authors [15]. The sRD-SIFT provides the best trade-off between performance and computational time, running at ≈ 0.97 fps. Both SIFT and sRD-SIFT were implemented in C code, while the pSIFT algorithm timings were obtained using a C-Mex implementation provided by the authors [15].

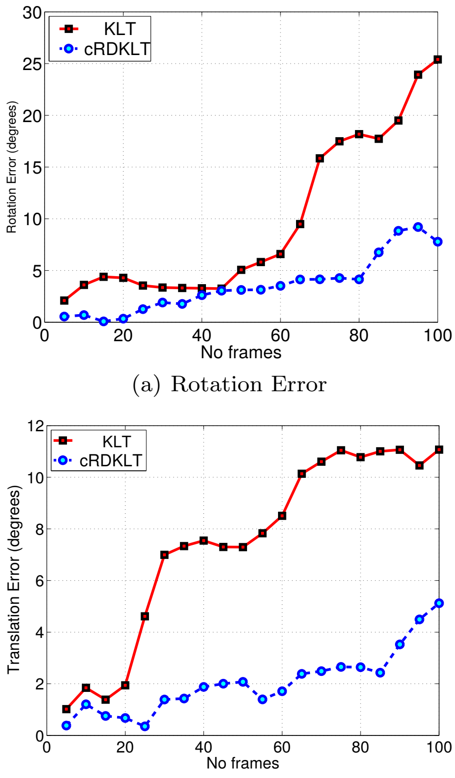
4.2 **VO** experiments

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 150 local images features, with feature replacement whenever a feature is lost.

Figure 6 shows that the motion estimation results. It can be observed that the cRD-KLT tracker provides better motion estimative in the sequential SfM pipeline meaning that the extra parameter in the RD compensated motion models permits a better convergence of the registration process in images presenting significant amounts of distortion.

Both methods were implemented in MATLAB/MEX files, with C-Mex files including only operations transver-

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



(a) Rotation Error

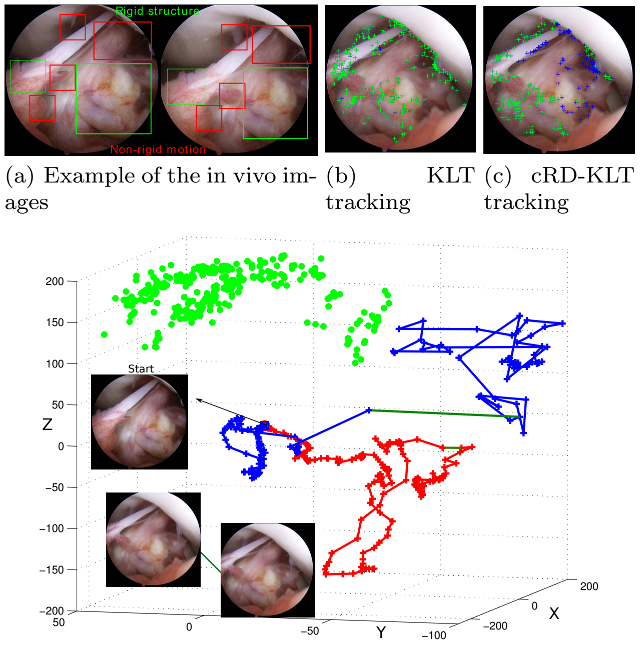
(b) Translation Angle error

Fig. 6 Visual odometry evaluation. The graphics show the rotation error 6(a) and translation error 6(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.

sal to both algorithms, such as interpolation routines, image gradient computation and image pyramid building. We have observed that KLT algorithm runs at ≈ 4.51 fps, while the cRD-KLT run at ≈ 4.34 fps. The small difference is probably due to the slightly more complex image model adopted in cRD-KLT [17], which requires a few more computations during the template tracking process.

4.3 *In vivo* experiments

In this experiment we evaluate the KLT and cRD-KLT trackers in a visual odometry experiment using orthopaedic in vivo data. We initialize the trackers with the same 300 local images features, with feature replacement whenever a feature is lost. 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.



(a) Example of the in vivo images (b) KLT tracking (c) cRD-KLT tracking

(d) Camera motion estimation

Fig. 7 Figure 7(a) shows two frames of the in vivo video sequences annotated with parts having rigid and non-rigid motion. Figure 7(b) 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). Figure 7(d) shows the recover motion 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 7 shows an example of the tracking results obtained with the KLT and cRD-KLT trackers. The final camera trajectory can be seen in Figure 7(d). 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.

5 Conclusions and final remarks

Structure-from-motion application in MIS is of vital importance for aiding the surgeon during navigation. Up until now, several studies have focused on such problem, by proposing robust estimation techniques and 2D-3D registration pipelines. In this paper, we analysed the role of the feature matching algorithms in which such methods rely. We have observed that due the low-texture and radial distortion effect arising in medical imagery, the SIFT and KLT algorithm are not the most viable options. The sRD-SIFT and cRD-KLT partially solve the problem of feature association in medical images, however such methods required the scene to have some texture variation. It was also observed that by improving feature tracking, reliable camera motion trajectories can be obtained in environments that combine rigid and non-rigid structures. Nevertheless, an open issue is the feature association in purely non-rigid environments. A topic we intend to investigate as future work.

Acknowledgements. Miguel Lourenço and João Barreto want to thank QREN-Mais Centro by generous funding through Novas Tecnologias para apoio à Saúde e Qualidade de Vida, Projecto A- Cirurgia e Diagnóstico Assistido por Computador Usando Imagem and the Portuguese Science Foundation by funding through grant SFRH/BD/63118/2009. Danail Stoyanov wants to acknowledge The Royal Academy of Engineerring Researh Fellowship for supporting his work.

Conflicts of interest. None.

References

1. Maximilian Allan, Steve S Thompson, Matthew Clarkson, Sebastien Ourselin, David J Hawkes, John Kelly, and Danail Stoyanov. 2d-3d pose tracking of rigid instruments in minimally invasive surgery. In *International Conference on Information Processing in Computer-Assisted Interventions*, 2014.
2. Simon Baker and Ian Matthews. Lucas-kanade 20 years on: A unifying framework'. *IJCV*, 56(3):221 – 255, March 2004.
3. Joao P. Barreto, Jose Roquette, Peter Sturm, and Fernando Fonseca. Automatic Camera Calibration Applied to Medical Endoscopy. In *BMVC*, 2009.
4. Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *CVIU*, 110(3):346–359, 2008.
5. Jean-Yves Bouguet. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm, 2000.
6. Darius Burschka, Ming Li, Masaru Ishii A, Russell H. Taylor, and Gregory D. Hager B. Scale-invariant registration of monocular endoscopic images to ct-scans for sinus surgery. In *MICCAI*, pages 413–421, 2004.
7. Ping-Lin Chang, Ankur Handa, Andrew Davison, Danail Stoyanov, and Philip Eddie; Edwards. Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking. In *International Conference on Information Processing in Computer-Assisted Interventions*, 2014.
8. Ping-Lin Chang, Danail Stoyanov, Andrew J. Davison, and PhilipEddie Edwards. Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2013*, volume 8149 of *Lecture Notes in Computer Science*, pages 42–49. Springer Berlin Heidelberg, 2013.
9. Toby Collins and Adrien Bartoli. 3d reconstruction in laparoscopy with close-range photometric stereo. In Nicholas Ayache, Herv Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, volume 7511 of *Lecture Notes in Computer Science*, pages 634–642. Springer Berlin Heidelberg, 2012.
10. Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24, 1981.
11. A.W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *IEEE-CVPR*, volume 1, pages I–125–I–132 vol.1.
12. Cristina García Cifuentes, Marc Sturzel, Frédéric Jurie, and Gabriel J. Brostow. Motion models that only work sometimes. In *BMVC*, 2012.
13. R. Garg, A. Roussos, and L. Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1272–1279, June 2013.
14. S. Giannarou, M. Visentini-Scarzanella, and Guang-Zhong Yang. Probabilistic tracking of affine-invariant anisotropic regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):130–143, 2013.
15. Peter Hansen, Peter Corke, and Wageeh Boles. Wide-Angle Visual Feature Matching for Outdoor Localization. *IJRR*, 29, 2010.
16. Tony Lindeberg. Feature Detection with Automatic Scale Selection. *Int. J. Comput. Vision*, 30, 1998.
17. M. Lourenco and J.P. Barreto. Tracking feature points in uncalibrated images with radial distortion. In *ECCV*, pages 752 –760, 2012.
18. M. Lourenco, J.P. Barreto, and F. Vasconcelos. sRD-SIFT: Keypoint Detection and Matching in Images With Radial Distortion. *IEEE-TRO*, 28(3):752 –760, june 2012.
19. Miguel Lourenco, Danail Stoyanov, and Joao P Barreto. Visual odometry in stereo endoscopy by using pearl to handle partial scene deformation. In *Medical Image Computing and Computer Assisted Intervention*. Springer Berlin Heidelberg, 2014.
20. David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60, 2004.
21. Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, 1981.
22. Xiongbiao Luo and K. Mori. Robust endoscope motion estimation via an animated particle filter for electromagnetically navigated endoscopy. *Biomedical Engineering, IEEE Transactions on*, 61(1):85–95, Jan 2014.
23. Y. Ma, S. Soatto, J. Kosecka, and S. Sastry. *An Invitation to 3D Vision: From Images to Geometric Models*. Springer-Verlag, 2003.
24. 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*, 17(8):974 – 996, 2013.

25. A. Malti and J.P. Barreto. Robust hand-eye calibration for computer aided medical endoscopy. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 5543–5549, May 2010.
26. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65:2005, 2005.
27. 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-TMI*, 31(4), 2012.
28. Maher Moakher. Means and averaging in the group of rotations. *SIAM J. Matrix Anal. Appl.*, 24(1):1–16, January 2002.
29. P. Mountney, D. Stoyanov, and Guang-Zhong Yang. Three-dimensional tissue deformation recovery and tracking. *IEEE-SPM*, 27(4), july 2010.
30. Peter Mountney and Guang-Zhong Yang. Motion compensated slam for image guided surgery. In Tianzi Jiang, Nassir Navab, JosienP.W. Pluim, and MaxA. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2010*, volume 6362 of *Lecture Notes in Computer Science*, pages 496–504. Springer Berlin Heidelberg, 2010.
31. Peter Mountney and Guang-Zhong Yang. Motion compensated slam for image guided surgery. In *MICCAI*, MICCAI'10, pages 496–504, 2010.
32. David Nistér. An Efficient Solution to the Five-Point Relative Pose Problem. *IEEE-TPAMI*, 26, 2004.
33. G. A. Puerto-Souza, A. Staranowicz, C. Bell, P. Valdastri, and G.L. Mariottini. "a comparative study of egomotion estimation algorithms for teleoperated robotic endoscope. In *CARE Workshop - (MICCAI'14)*. Springer Berlin Heidelberg, 2014.
34. Gustavo A. Puerto Souza and Gian Luca Mariottini. A comparative study of correspondence-search algorithms in mis images. In *MICCAI*, pages 625–633, 2012.
35. Gustavo A. Puerto Souza and Gian Luca Mariottini. Hierarchical multi-affine (hma) algorithm for fast and accurate feature matching in minimally-invasive surgical images. In *IEEE-IROS*, pages 2007–2012, 2012.
36. Danail Stoyanov. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In Nicholas Ayache, Herv Delingette, Polina Golland, and Kensaku Mori, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2012*, volume 7510 of *Lecture Notes in Computer Science*, pages 479–486. Springer Berlin Heidelberg, 2012.
37. Tina Y. Tian, Carlo Tomasi, and David J. Heeger. Comparison of Approaches to Egomotion Computation. In *IEEE-CVPR*, 1996.
38. H. Wang, D Mirota, M. Ishii, and G. D. Hager. Robust motion estimation and structure recovery from endoscopic image sequences with an adaptive scale kernel consensus estimator. In *IEEE-CVPR*, pages 1–7, June 2008.
39. M. C. Yip, D. G. Lowe, S. E. Salcudean, R. N. Rohling, and C. Y. Nguan. Tissue tracking and registration for image-guided surgery. *IEEE-TMI*, 31(11):2169–2182, Nov.