Evaluation of Keypoint Detectors and Descriptors in Arthroscopic Images for Feature-Based Matching Applications

Andres Marmol, Thierry Peynot, Anders Eriksson, Anjali Jaiprakash, Jonathan Roberts, and Ross Crawford

***Abstract*—Knee arthroscopy is a very challenging surgical procedure that would strongly benefit from systems that can continuously map the inside of the knee, localize the arthroscope and surgical tools, and control instruments using visual information. A fundamental requirement of most of these systems is the correct and fast matching of visual features. Feature-based systems have been demonstrated in laparoscopy but have yet to be extended to the context of arthroscopy. As an essential initial step, this letter proposes the first detailed experimental evaluation of the performance of state-of-the-art feature detection and description methods on arthroscopic images. We first evaluate the behavior of eight keypoint detectors under 133 setting combinations using four different metrics in a dataset with 100 *in-vivo* images. We then combine the previous detectors with six feature descriptors and evaluate the matching performance for the resulting features (detector+descriptor) across five different image transformations. A validation is performed using *in-vivo* images acquired under varying camera motion and illumination. The results show that the best-performing feature in knee-arthroscopy images is DoG+SIFT, while features BRISK+SURF and BRISK+BRISK are recommended for viable implementations in real time.**

***Index Terms*—Arthroscopy, computer vision for medical robotics, feature-based SLAM, medical robots and systems, visual navigation.**

I. Introduction

KNEE arthroscopy is the most common minimally invasive orthopedic procedure in the U.S. [1] and the world. During this procedure, a camera and an arthroscope allow surgeons to observe unstructured and narrow views of the inside of the knee as illustrated in Fig. 1. Given such visually challenging monocular images, the surgeon needs to a) estimate where the camera and the instruments are within the knee, b) maintain a mental map of the knee environment, and c) perform the appropriate therapeutic action while manipulating multiple instruments. These tasks are both mentally and physically demanding for surgeons [2]. The objective of our research is to develop robotic vision systems that can a) provide an accurate 3D map of the inside of the knee, b) localize the arthroscope and other surgical tools, and c) perform various surgical actions under the surgeon's supervision. We believe these systems to be an essential step towards the development of a semi-autonomous robotic assistant for orthopedic surgery.

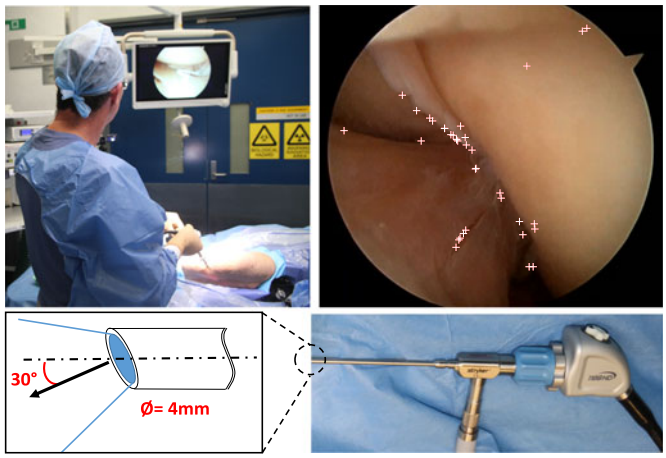

Fig. 1. Arthroscopy procedure and equipment. Top-left: Surgeon ambidextrously manipulating surgical instruments with his attention directed to a display of the arthroscope images. Top-right: Image captured inside the knee with features overlaid (white crosses). Bottom: Camera head and arthroscope with detail of the arthroscope's 30° oblique optics. Images better seen in color.

Techniques such as Structure-from-Motion (SfM), visual Simultaneous Localisation and Mapping (SLAM) and visual servoing have reached maturity in recent years and are becoming fundamental components of surgical robots [3], [4]. Tissues are typically imaged passively through small incisions in the skin. Although active sensing systems (e.g. structured light) have been employed in Minimally Invasive Surgery (MIS), their performance is typically limited by the medium's attenuation and the device miniaturization [5]. Besides their own specific requirements, these techniques require an underlying visual tracking system that is robust, accurate and fast to compute [6]. Common tracking systems include dense and feature-based approaches, with the latter group shown to be more accurate [7]. Consequently, we focus this letter on visual feature-based matching applications.

We refer to a feature as the mathematical representation of a distinct image pattern, resulting from consecutive processes of

Manuscript received February 10, 2017; accepted May 24, 2017. Date of publication June 9, 2017; date of current version July 19, 2017. This letter was recommended for publication by Associate Editor A. Krupa and Editor F. Chaumette upon evaluation of the reviewers' comments. This work was supported by the Australian Centre for Robotic Vision. *(Corresponding author: Andres Marmol.)*

A. Marmol, T. Peynot, A. Eriksson, A. Jaiprakash, and J. Roberts are with the Australian Centre for Robotic Vision, Queensland University of Technology, Brisbane, 4000 QLD, Australia (e-mail: andres.marmolvelez@qut.edu.au; t.peynot@qut.edu.au; anders.eriksson@qut.edu.au; anjali.jaiprakash@qut.edu.au; jonathan.roberts@qut.edu.au).

R. Crawford is with the Australian Centre for Robotic Vision, Queensland University of Technology, Brisbane, 4000 QLD, Australia, and also with the Prince Charles Hospital, Brisbane, 4032 QLD, Australia (e-mail: r.crawford@qut.edu.au).

Digital Object Identifier 10.1109/LRA.2017.2714150

detection and *description*; therefore, we label them by *detector + descriptor*. In the medical context, feature matching–based applications tailored for laparoscopy have reported promising results using monocular images [4], [8]–[10]. However, to the best of authors' knowledge, no similar system has been demonstrated in arthroscopy.

Arthroscopic images portray tissues such as cartilage and ligaments whose appearance and texture greatly differs from what can be observed in the abdominal area. When facing a new visual context such as in arthroscopy, it is essential to thoroughly evaluate the performance of state-of-the-art feature detection and description methods on relevant images. Consequently, we propose the first detailed evaluation of detection and description methods in arthroscopic images and suggests the most appropriate features for matching.

Two studies are presented in this letter. In Study 1, we evaluate the repeatability of 8 keypoint detectors under 133 settings combinations with a dataset of 100 arthroscopic images. This provides a baseline set of parameter values that are appropriate for keypoint detection in the context of knee arthroscopy. Study 2 combines the previous detectors with 6 feature descriptors and evaluates the features' matching rates across 5 different synthetic image transformations. Features with the best performance are further validated under real imaging conditions. Overall, the best performance was obtained with feature DoG+SIFT. Features BRISK+SURF and BRISK+BRISK showed slightly reduced performance but with significantly faster computation speeds. By demonstrating which features are most likely to allow for a successful implementation of matching-based techniques in arthroscopy, this work constitutes an essential step towards the development of robotic assistants for orthopedic surgery.

II. Literature Review

Feature detection and description have been central topics of the computer vision community for nearly 20 years. In detection, keypoints such as corners are identified at locations where image properties clearly differ from the surrounding neighborhood [11]. In description, a keypoint's neighborhood is used to abstract information that uniquely identifies that keypoint on different views. Robust invariance of this description is required for the establishment of reliable keypoint matches or correspondences, between different images [6]. While attempts at ensuring invariance had led to the implementation of computationally expensive floating-point descriptors, recent interest in real-time applications has motivated the use of compact binary descriptors.

Mikolajczyk *et al.* [12], [13] evaluated up to six detectors and ten descriptors in a dataset containing both structured and textured images under different imaging conditions. This evaluation framework has been widely accepted and used, e.g., in the evaluation of features for visual tracking [6]. Although the datasets used in those studies include different imaging conditions, they were limited to outdoor or man-made environments. Biological environments as observed in arthroscopy are smooth and highly unstructured, and hence, require a specific evaluation. Our performance evaluation is closer to the one presented in [6], with an emphasis on the overall performance of detector+descriptor combinations. We also propose to use a performance metric that is well suited for SLAM applications, such as the spatial spread of the features in the image.

The work by Mountney *et al.* [14] can be regarded as the first descriptor study with MIS images. Early studies compared the performance of a limited number of detectors and descriptors [15], [16]. Most recently, out-of-the-shelf features have been applied in endoscopic tracking applications [10], [17]. Nonetheless, the full potential of state-of-the-art features in endoscopy remains mostly unexplored [18].

To the best of our knowledge, *Arthronav* is the only project that has analyzed a feature's performance in arthroscopic images [19]. However, this study focused on robustifying SIFT features against strong radial distortion. Furthermore, the image dataset lacked realistic intraoperative conditions such as the presence of fluid and floating tissue.

Notably, most aforementioned studies used fixed values for the algorithms' parameters, typically the default values suggested by the original developers of each detection and description method. However, as demonstrated in [6], the performance of each method strongly depends on the values of each parameter and the type of environment captured in the images (e.g. urban environment vs. biological tissues). Therefore, the feature performance study proposed in this letter includes a thorough investigation of performance with respect to parameter values, and determines what settings are appropriate for arthroscopic images. Table I in Section III summarizes the proposed adjusted parameter values.

Most of these feature studies relied on software provided directly by the respective authors. Open source computer vision libraries such as BOOFCV [20], VLFeat [21] and OpenCV [22] have been introduced to facilitate the benchmarking of algorithms. We used OpenCV in this work.

III. Algorithms

This section describes the algorithms used for the detection and description of features in our studies. These include algorithms used previously in MIS footage as well as state-of-the-art ones not used in that context yet. The detection methods include three blob detectors, four corner detectors and a region detector. Table I summarizes the eight detectors and their adjustable parameters. The description methods comprise two floating-point as well as four binary descriptors.

A. Keypoint Detectors

1) Difference of Gaussians (DoG): The DoG is a scale-invariant detector used for blob extraction based on the Laplacian of an image [23].

2) Determinant of Hessian (DoH): The DoH is a scale-invariant detector used for blob extraction based only on the determinant of a simplified Hessian matrix of image intensities [24].

3) Center Surrounded Extrema (CenSurE): This method was proposed for computing accurate large-scale features by applying consecutive center-surround filters [25].

4) Features from Accelerated Segment Test (FAST)-Based detectors: Features from Accelerated Segment Test (FAST)-based detectors were originally inspired by morphological corner detectors. In *FAST* the central pixel of a Bresenham circle is

TABLE I
EIGHT DETECTORS ALONG WITH THEIR ADJUSTABLE PARAMETERS

DoG	Type	#Val	#Cmb.	**MSER**	Type	#Val	#Cmb.	**CenSurE**	Type	#Val	#Cmb.
NOctaveLayers(3)	*int*	8		Delta(2)	*int*	5		ResponseThreshold(5)	*int*	8	
ConstrastThreshold(0.01)	*float*	7		MinArea(5)	*int*	6		MaxSize(45)	*int*	8	
EdgeThreshold(10)	*int*	5	**23**	MaxArea(300000)	*int*	5	**18**	LineThresholdProjected(10)	*int*	5	**27**
Sigma(1.6)	*float*	6		MaxVariation(0.2)	*float*	5		LineThresholdBinarized(8)	*int*	5	
								SuppressNonmaxSize(6)	*int*	5	
DoH	Type	#Val.	#Cmb.	**BRISK**	Type	#Val.	#Cmb.	**oFAST**	Type	#Val.	#Cmb.
HessianThreshold(55)	*int*	7		Threshold(17)	*int*	6		ScaleFactor(1.3)	*float*	5	
NOctaves(4)	*int*	6	**17**	Octaves(3)	*int*	5	**14**	NLevels(8)	*int*	6	**22**
NOctaveLayers(3)	*int*	6		PatternScale(1.0)	*float*	5		EdgeThreshold(30)	*int*	5	
								FastThreshold(16)	*int*	6	
FAST	Type	#Val.	#Cmb.	**AGAST**	Type	#Val.	#Cmb.				
Threshold(12)	*int*	5	**7**	Threshold(10)	*int*	5	**8**				
Type('TYPE_9_16')	*enum*	3		Type('OAST_9_16')	*enum*	4					

used to classify surrounding pixels and propose keypoint candidates [26].

The *The Adaptive and Generic Accelerated Segment Test (AGAST)* detector builds upon the previous method and avoids retraining in new environments by means of an adaptive algorithm [27]. The *Invariant Scalable Keypoints (BRISK)* detector was proposed later to obtain FAST scale-invariant keypoints [28]. Lastly, the *Oriented FAST (oFAST)* detector included efficient scale and orientation estimation [29].

5) Maximally Stable Extremal Regions (MSER): This region detector identifies connected components over a series of thresholded images [30].

B. Feature Descriptors

1) Scale-Invariant Feature Transform (SIFT): The SIFT algorithm encodes the gradient and orientation around keypoints using a 128-byte vector [23].

2) Speeded Up Robust Features (SURF): In this algorithm Haar wavelets and integral images are efficiently combined to produce a 64-byte feature vector [24].

3) Binary Robust Independent Elementary Features (BRIEF): The BRIEF descriptor reduces the computation time of the previous floating-point descriptors using a binarized 32-byte vector [31].

4) Rotated BRIEF (rBRIEF): An angle-steered version of the previous BRIEF descriptor, it exploits rotation information of, typically, an oFAST keypoint [29].

5) BRISK: Another improvement over BRIEF is achieved by normalizing keypoints' orientation to give rotation invariance [28]. A 64-byte string is used for description.

6) Fast Retina Keypoint (FREAK): This descriptor employs a circular sampling grid to efficiently encode image information into a 64-bytes binary string [32].

IV. EXPERIMENTAL METHOD

As shown in the previous section, the implementations of keypoint detectors comprise numerous parameters that need to be tuned appropriately for the relevant context, i.e. arthroscopy in this work. Therefore, in a first study we evaluate the detector's repeatability under varying parameter values and identify relevant settings for arthroscopic images. Then, given these settings, a second study evaluates the matching performance of detector+descriptor methods. Following a final validation under real imaging conditions, we recommend the features best suited for matching arthroscopic images.

A. Study 1: Settings' Influence on the Detectors' Performance

In Study 1, the values of selected settings on each of the 8 keypoint detectors were initially varied over a broad sweep range. Following this *coarse* run, new default values were selected to detect at least 100 keypoints across images in the dataset. This allowed for the visualization of a trend in the metrics during a parameter sweep. As an example, DoG's `ContrastThreshold` default value of 0.04 (as given in OpenCV) was redefined to 0.01. Following a similar procedure for all other detectors, new default values and sweep ranges were defined. To keep the number of combinations tractable during the study, settings not being varied were kept at their adjusted default values. Overall, Study 1 reported 4 metrics across 133 detectors' settings combinations.

1) Datasets: The *KneeRegions dataset* was composed by 100 color images of 1280 $\times$ 720 pixels which were processed as *uint8* arrays of gray-scale pixel intensities. The images were acquired with Stryker equipment[1] during a left knee arthroscopy on a cadaver. Data acquisition was approved by the Australian National Health and Medical Research Council (NHMRC) - Registered Committee Number EC00171 under Approval Number 1400000856.

Regions of interest were selected as instructed by an expert orthopedic surgeon and included: lateral gutter (10 images), medial gutter (10), patella and femoral trochlea (10) suprapatellar pouch (10), lateral compartment (20), medial compartment (20), and the anterior cruciate ligaments (20). Examples of these locations can be observed in Fig. 2. Each 10-image subset was extracted from a short video and therefore contained scenes with similar viewpoint. The latter three subsets included two dissimilar viewpoints.

The camera featured three 1/3" progressive CCD image sensors with 60 Hz refresh rate. Similar or higher resolution (1920 $\times$ 1080 px) can be found in Karl Storz, Richard Wolf, Arthrex, Olympus and ConMed camera heads, which cover most of the arthroscopy and laparoscopy imaging solutions in the market. Similarly, other commercial arthroscopes feature typically a diameter of 4 mm, a 30° direction of view and 115° wide angle field of view. While 0° or 70° angles of view are also available, the field of view and distortion remain similar. Consequently, we expect the conclusions of our studies to be generalizable to similar arthroscopic equipment.

In order to provide ground truth for the repeatability analysis, every image in the *KneeRegions dataset* was synthetically

[1]Stryker 1188 HD Autoclavable 3-Chip Camera, Stryker Arthroscope Ideal Eyes Ref.: 0502-704-030.

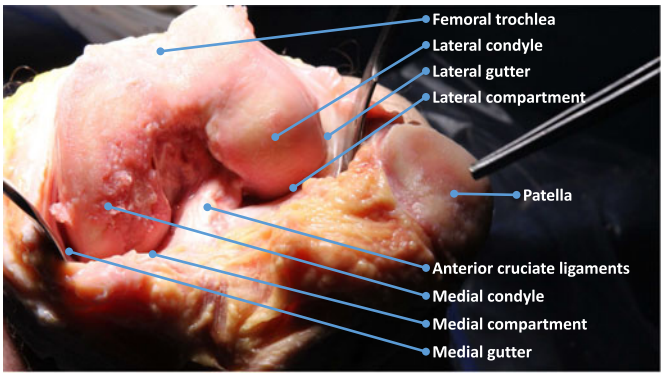

Fig. 2. Regions of diagnostic interest shown in an image of a human knee arthrotomy. Image better seen in color.

modified to produce one hundred 10-frame virtual sequences representing common camera movements. Camera motions that commonly arise during manipulation through the skin's insertion point (fulcrum effect) include pan/tilt, roll, insertion and retraction.

The magnitudes of the transformations were estimated by manipulating an arthroscope in a bendable wet knee phantom[2]. The arthroscope was inserted at a location that allowed analogous views as the ones observed during in-vivo arthroscopies (see Fig. 1). From this reference position, the camera was panned/tilted/rolled [−30°...30°] and inserted/retracted [−6 mm...+20 mm]. Outside the given ranges, motion was either not possible or the content overlap with respect to the reference image was inappropriately small.

The projective changes caused by the constrained camera motion were approximated as follows: Panning and tilting resembled a ≈330 pixel 2D translation. Rolling resembled an off-center image rotation of up to 30°. Retraction and insertion resembled image down and up scaling by 0.75 and 1.5 factors respectively. Consequently, the virtual sequences were constructed using affine transformations with constant relative increments. The final homography transformed the reference image by a ≈330 pixels 2D translation, a 30° rotation and a 0.75 scaling. The content overlap with respect to the reference image varied from 91.22% (frame 1) down to 34.38% (frame 10).

2) Metrics: **Repeatability**, as defined in Equation (1), is the ratio of the number of points repeated between two images I_1 and I_i to the total number of commonly detected points.

$$r_i(\epsilon) = \frac{|R_i(\epsilon)|}{\min(n_1, n_i)} \tag{1}$$

$$R_i(\epsilon) = \{(\widetilde{x}_1, \widetilde{x}_i) | dist(H_{i1}\widetilde{x}_1, \widetilde{x}_i < \epsilon)\}$$

$$\widetilde{x}_1 = \{x_1 | H_{1i}x_1 \in I_i\}, \qquad \widetilde{x}_i = \{x_i | H_{i1}x_i \in I_1\}$$

where x_1 and x_i are the keypoints in images I_1 and I_i respectively, n_1 and n_i are the number of keypoints in images I_1 and I_i respectively, and H_{ij} is the homography between images I_i and I_j. OpenCV's default $\epsilon = 1$ was used in our evaluation.

The metric, although widely accepted, is known to have two main limitations [6]. Firstly, an algorithm that "detects" every single pixel in an image would have $r_i(\epsilon) = 1$. We also report the number of keypoints to inform about this bias. Secondly, an algorithm detecting just one point in image I_i would achieve perfect repeatability if that sole point was detected previously. We do not expect this to occur as our choice of adjusted default values already sets a lower bound on the number of expected keypoints.

numKeypoints: number of keypoints detected in an image by a detection method.

Previous deployment of SLAM approaches in MIS ([8]–[10]) demonstrated that detection, description and matching could be executed at frame rates. Owing to properties of multiview-geometry, more accurate estimation of the motion experienced between two views can be performed if the features that are matched are also *well* distributed over the images. Therefore, additional metrics are considered to provide insights about the detectors' suitability for feature-based matching applications.

compTime: time in seconds taken by a detection method to find all keypoints in an image. Computation was conducted on a Windows 7 computer with an Intel core i7-4790 3.60 GHz CPU and 16 GB of RAM. *OpenCV 3.1.0* libraries were compiled for Matlab using *mexopencv*[3]. Recorded times are indicative only as no extensive profiling was conducted.

1-spread spatial distribution of the detected keypoints in an image. To quantify the spread, first a uniform 10x10 grid was overlaid on the image. Cells not completely inside the arthroscope's eyepiece were considered invalid (see Fig. 1). The 1-spread was then computed as the number of valid cells containing at least one keypoint divided by the total number of valid grid cells.

B. Study 2: Feature Matching Performance

The second study built upon the previous detectors' settings and combined them with 6 feature descriptors to evaluate the features' matching performance. Synthetically modified real images provided a ground truth for the evaluation of the percentage of *true positive correspondences*, *%TP*, between the *original* and the *transformed* images. Overall, Study 2 evaluated 798 detector+descriptor combinations using 2 main metrics across 500 image-pairs.

1) Datasets: The *synthTransf. dataset* included four subsets constructed using the *KneeRegions dataset* and the maximum expected projective changes due to camera motion. The subsets were labeled *Tx* (≈330 pixel shift), *Rx* (30° rotation), *Sc.Dw* (0.75x reduction), *Sc.Up* (1.5x magnification). A fifth subset, *Def*, emulated tissue deformation that takes place when the surgeon occasionally repositions the patient's leg to gain access to certain knee regions. We emulated such changes using a non-linear Gaussian Mixture Model (GMM) [33] with a maximum pixel shift of 50.

2) Metrics: As described previously, performance of feature-based applications can be linked to the correct association of a sufficient number of correspondences. We made use of the metric *%TP* to select the features with the highest matching performance.

%TP: the percentage of correctly matched keypoints out of the total number of matches, also referred to as *precision* in [6]. Correspondences were matched using the Sum of Squared Differences (SSD) (floating-point descriptors) or the Hamming distance (binary descriptors). The known image transformations in the *synthTransf. dataset* were used to compute the expected

[2]Sawbones wet arthroscopy knee ERP #1400.

[3]https://kyamagu.github.io/mexopencv/

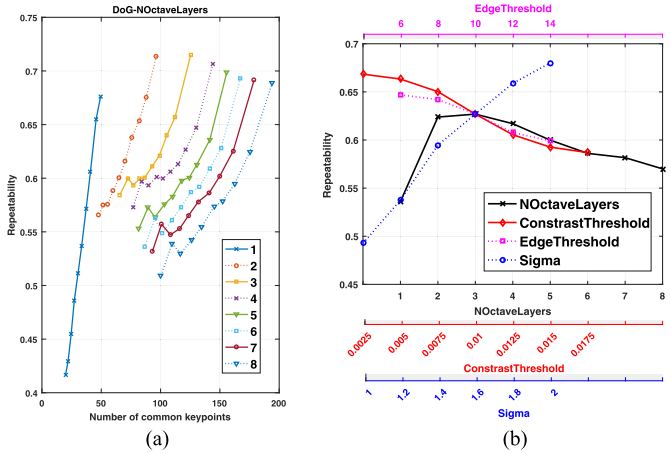

Fig. 3. DoG results per adjustable parameter: (a) Repeatability versus number of common keypoints. Each trend-line corresponds to a different value of `NOctaveLayers`. (b) Average repeatability over the sequence for each parameter-value pairs. Figures better seen in color.

keypoint locations after the original image was warped. A match was considered correct if the corresponding keypoint in the warped image laid in a two-pixel neighborhood from the expected ground-truth location.

1-spread: spatial distribution of the correct matches, computed as described in Study 1. This metric was also used to avoid misleading performance results (e.g. a single correct match that has %TP = 100%).

V. Experimental Results

A. Study 1: Settings' Influence on the Detectors' Performance

The metrics *repeatability*, *numKeypoints*, *1-spread* and *compTime* were used to assess the impact of each individual parameter on the detection performance. Parameters were changed between values a and b, and, for each value, the metrics were computed in all of the 100 images in the *KneeRegions* dataset. The median of a metric in the dataset was reported for the sweeping boundaries a and b in the format $\{metric\_value@a...metric\_value@b\}$.

1) DoG: Fig. 3 illustrates the repeatability rates summarized in Table II. The repeatability between the reference and the 10 virtual images was computed for each of the selected parameter' settings, e.g., `NOctaveLayers` in Fig. 3(a). The number of common keypoints between the image pairs was also recorded. The repeatability rates vs. the number of common keypoints were computed as an average over the 100 images in the *KneeRegions dataset*. The metrics decrease (from the upper right corner towards the bottom left corner) as the reference and virtual images are further apart in the sequence.

Fig. 3(b) synthesizes the results for all adjustable parameters of the DoG detector by averaging the repeatability over the sequence. Settings `contrastThreshold` and `edgeThreshold` did not influence significantly the repeatability but could be used to alter *numKeypoints*. Repeatability for parameter `NOctaveLayers` was consistent for two or more layers, although there was a noticeable performance drop at setting value 1. The parameter `sigma` resulted in the largest repeatability variation. Larger setting values led to better repeatability at the expense of *numKeypoints*. At the adjusted default values, shown as a common intersection point in Fig. 3(b), 162 keypoints, with 48% spread and 63% repeatability were detected in about 0.31 second.

2) DoH: Parameters `NOctaves` and `NOctaveLayers` did not influence significantly the repeatability but could be used to alter *numKeypoints*. Larger setting values were observed to increase the computation time of up to 20% in both parameters. Small `hessianThreshold` values increased the repeatability but the large number of keypoints detected (≈2000) is likely to bias this metric. At the adjusted default values, 225 keypoints with 49% spread and 48% repeatability were detected in about 0.09 second.

3) CenSurE: Parameters `responseThreshold`, `suppressNonMaxSize` and `LineThresholdBinarized` did not influence significantly the repeatability but could be used to alter *numKeypoints*. An exception took place at `LineThresholdBinarized`$\leq$4, where all keypoints were filtered out. Settings `LineThresholdProjected`=10 and `maxSize`=32 resulted in the maximum repeatability at each respective parameter. At the adjusted default values, 75 keypoints with 30% spread and 64% repeatability were detected in about 0.03 second.

4) FAST-Based Methods: A reduction of the main `threshold` parameter in all FAST-based methods resulted in an exponential increase in *numKeypoints*. Consequently, repeatability results are to be used with caution due the metric inherent bias. The computation time did not exceed 0.02 second across all combinations tested. This is at least one order of magnitude faster than DoG. In *FAST* and *AGAST* detectors, 7–12 and 9–16 patterns resulted in approximately the same performance, while pattern 5–8 could lead to insufficient detection. For *BRISK*, parameters `Octaves` and `PatternScale` did not influence significantly the repeatability but could be used to alter *numKeypoints*. Similarly, parameters `ScaleFactor`, `NLevels` and `EdgeThreshold` had negligible influence on the *oFAST* repeatability.

5) MSER: Parameter `maxVariation` did not influence significantly the repeatability but could be used to alter *numKeypoints*. The computation time strongly depended on the filtering made with parameters `minArea` and `maxArea`. The detector was biased towards finding large area features. This could be attributed to the tissue smoothness observed in the images. However, allowing the computation of these large areas led to significant computation times, e.g. a couple of seconds for areas larger than 300.000 pixels. The parameter `delta` had a maximum repeatability at setting value 2, dropping rapidly for settings larger that 4.

All of the detectors' repeatability rates were consistent across the 10 regions of diagnostic interest except for MSER (not shown). Overall FAST, oFAST and AGAST performed similarly with about 65% averaged repeatability, $\overline{r}(1)$. DoG, BRISK and CenSurE followed with $\overline{r}(1) = 60\%$, while DoH and MSER ranked lastly with $\overline{r}(1) = 48\%$ and $\overline{r}(1) = 37\%$, respectively.

In summary, Study 1 has established meaningful default values and tuning ranges for the detectors' settings using assorted knee arthroscopy images. Several settings have been found not to influence the detection while others have led to insufficient detection or impractical processing times in the context of real-time feature matching. In particular, results show that MSER is not a reliable detector for arthroscopy.

TABLE II
DETECTORS' PERFORMANCE: *repeatability*, *numKeypoints*, AND *1-spread* METRICS FOR EACH DETECTOR PARAMETERS (ROUNDED HALF UP)

DoG	*Repeatability(%), numKeypoints, 1-spread(%)*	**MSER**	*Repeatability(%), numKeypoints, 1-spread(%)*
NOctaveLay. (1)	{54%@1...57%@8}, {73...269}, {37%...55%}	Delta (1)	{37%@1...14%@5}, {313...61}, {76%...5%}
Const.Thr. (0.025)	{67%@0.0025...59%@0.0175}, {1000...48}, {98%...25%}	MaxArea (≈3e5)	{5%@14400...38%@1.2e6}, {23...187}, {13%...22%}
EdgeThr. (2)	{65%@6...60%@14}, {144...170}, {46%...48%}	MinArea (23)	{40%@5...15%@120}, {107...56}, {19%...10%}
Sigma (0.2)	{49%@1...68%@2}, {484...122}, {72%...45%}	MaxVar. (0.2)	{45%@0.2...42%@1}, {107...403}, {19%...75%}
DoH	*Repeatability(%), numKeypoints, 1-spread(%)*	**BRISK**	*Repeatability(%), numKeypoints, 1-spread(%)*
HessianThr. (15)	{61%@10...46%@100}, {1994...97}, {98%...33%}	Threshold (2)	{65%@11...56%@21}, {856...36}, {84%...24%}
NOctaves (1)	{50%@2...48%@7}, {178...229}, {43%...50%}	Octaves (1)	{59%@1...56%@5}, {117...137}, {46%...47%}
NOctaveLay. (1)	{49%@2...47%@7}, {177...364}, {47%...54%}	Patt.Sc. (0.2)	{57%@0.6...57%@1.4}, {134...134}, {47%...46%}
CenSurE	*Repeatability(%), numKeypoints, 1-spread(%)*	**oFAST**	*Repeatability(%), numKeypoints, 1-spread(%)*
ResponseThr. (1)	{66%@1...60%@8}, {438...25}, {76%...14%}	Sc.Fact. (0.2)	{64%@1.1...64%@1.9}, {147...147}, {51%...51%}
MaxSize (var)	{61%@8...34%@128}, {58...18}, {29%...12%}	NLevels (1)	{64%@4...64%@9}, {147...147}, {51%...51%}
LineThrProj. (5)	{51%@5...56%@25}, {40...89}, {22%...33%}	EdgeThr. (15)	{64%@15...63%@75}, {148...145}, {52%...47%}
LineThrBin. (4)	{0%@4...63%@20}, {0...87}, {0%...34%}	FastThr. (2)	{73%@8...60%@18}, {1315...82}, {84%...38%}
SupNonMax. (2)	{64%@2...64%@10}, {153...56}, {37%...27%}		
FAST	*Repeatability(%), numKeypoints, 1-spread(%)*	**AGAST**	*Repeatability(%), numKeypoints, 1-spread(%)*
Threshold (4)	{88%@4...60%@20}, {6271...44}, {99%...28%}	Threshold (2)	{68%@6...60%@14}, {3684...296}, {98%...66%}
Type (var)	{0%@5_8...71%@9_16}, {21...500}, {16%...76%}	Type (var)	{63%@5_8...65%@9_16}, {575...1016}, {79%..87%}

Metrics are reported for both sweeping boundaries, but are only explicitly stated for the repeatability metric. Numbers in parenthesis indicate the uniform step size used during parameter variation, e.g. DoG's `NOctaveLayers` is changed from 1 to 8 in increments of 1. Variable (var) step size is reported for parameters swept with non-uniform step size due to OpenCV implementation particularities.

B. Study 2: Feature Matching Performance

The matching performance results of Study 2 are summarized in Table III. The table is divided in six blocks for each descriptor. The 8 detectors are arranged row-wise. The 5 synthetic transformations are categorized column-wise. The performance of each of the features in each of the *synthTransf. subsets* are scored according to the *%TP*. The categories include: 0–10%, 10–50%, 50–90%, 90–100% and *Inconclusive* for those methods whose performance varies across more than two categories when using different settings. Features in this latter category are considered particularly fragile. Borderline performance between two categories is indicated with either ↑ or ↓. An asterisk indicates that the resulting matches have less than 10% *1-spread*.

Considering the general performance of each descriptor, BRIEF, rBRIEF and FREAK seemed unsuited for correspondence matching of arthroscopic images. Correct matches were rarely established and when done, their spread was always below 10%.

As observed in Table III, only three features (FAST+SURF, CenSurE+SURF and CenSurE+BRISK) were able to match correspondences across all transformations with *%TP*>50%. Nonetheless, performance was fragile and decayed strongly after small settings' changes. Considering only affine transformations and criteria of *%TP*>50% and *1-spread*>10% resulted in the best-suited candidates for feature-based matching techniques being: DoG+SIFT, DoG+SURF, BRISK+SURF and BRISK+BRISK. These features performed poorly in images undergoing the nonlinear transformation *Def.*. Consequently, we proceed to validate the results of Study 2 assuming that the patient's leg is not repositioned. We propose to evaluate the effects of real limb reconfiguration in future work.

C. Feature Matching Performance Validation

The best ranking features from Study 2 were validated according to their matching performance and number of matches using unmodified pairs of images from arthroscopic videos.

A *KneeValidation dataset* was constructed from 10 subsets containing 6 images each. Subsets *A-D* were acquired at 4 different areas of the knee with small camera motion, constant illumination and slight tissue deformation due to water flow. An expert estimated the maximum pixel shift, *t*, across all images to be 20. Similarly, subsets *E-H* captured 4 different areas of the knee under larger camera motion ($t = 60$). The remainder two subsets displayed illumination changes in 2 areas of the knee with both a static and a moving camera ($t = 5$ and $t = 55$ respectively). In order to change the illumination conditions, the brightness setting of a commercial fiber optic light source[4] was varied in uniform increments over the full adjustable range.

A RANSAC-based homography was used to exclude outliers before validation. The correctness of a match was decided by comparing the euclidean distance ($dist$) between correspondences against the maximum expected image shift of each dataset. Due to the lack of ground truth, it must be noted that pairs for which $dist < t$ holds are not *necessarily* correct. Similarly, although we illustrate the outlier removal process with the well-known RANSAC method, better estimation methods such as MPSAC [34] could be employed to avoid undesired removal of stable inliers.

The validation results are summarized in Table IV. DoG+SIFT features were found to be both correct and numerous. BRISK+BRISK and BRISK+SURF features were also correctly matched albeit in fewer numbers. Lastly, DoG+SURF features established more matches than the previous two methods, but with much lower precision.

It is worth mentioning that the impact of external illumination is much less pronounced than in other MIS procedures. This is to be expected in arthroscopy where the tissues are scarcely a few millimeters in front of both camera and light source. We verify this by comparing subsets with and without varying lighting conditions (A-D vs. I and E-H vs. L). Although the light

[4]Stryker L9000 LED Light Source.

TABLE III
FEATURE MATCHING RESULTS IN THE *synthTransf. dataset*: PERFORMANCE OF THE EIGHT DETECTORS WHEN COMBINED WITH EACH OF THE SIX DESCRIPTORS

	SIFT descriptor					SURF descriptor					BRIEF descriptor				
	Tx	Rx	Sc.Dw	Sc.Up	Def	Tx	Rx	Sc.Dw	Sc.Up	Def	Tx	Rx	Sc.Dw	Sc.Up	Def
G	90-100	90-100	50-90↑	50-90↑	0-10*	90-100	50-90↑	50-90	50-90	0-10*	50-90↑*	0-10*	50-90↓*	0-10*	0-10*
H	90-100*	10-50	50-90	50-90	0-10*	90-100	10-50	50-90	10-50	0-10*	90-100↓*	0-10*	10-50	0-10*	0-10*
C	90-100*	0-10*	50-90	50-90↓*	50-90*	90-100*	50-90↓	50-90	10-50↑*	10-50↑*	90-100*	0-10*	50-90↓*	10-50*	50-90
F	90-100	0-10*	50-90	0-10*	50-90	90-100	50-90	50-90↑	10-50↑*	50-90	Inc.*	0-10*	Inc.*	0-10*	Inc.*
A	90-100	0-10*	50-90	10-50*	90-100	50-90	0-10*	10-50	10-50	0-10*	Inc.*	0-10*	Inc.*	0-10*	10-50↓*
B	90-100*	50-90	50-90	50-90*	0-10*	90-100	50-90	50-90	50-90↓	0-10*	90-100*	0-10*	50-90*	10-50*	0-10*
O	90-100*	50-90	10-50*	0-10*	10-50*	90-100*	50-90	50-90	10-50*	50-90	90-100	0-10*	50-90	10-50*	50-90
M	0-10*	0-10*	Inc.*	0-10*	0-10*	Inc.*	Inc.*	50-90*	Inc.*	0-10*	0-10*	0-10*	10-50*	0-10*	0-10*

	rBRIEF descriptor					BRISK descriptor					FREAK descriptor				
	Tx	Rx	Sc.Dw	Sc.Up	Def	Tx	Rx	Sc.Dw	Sc.Up	Def	Tx	Rx	Sc.Dw	Sc.Up	Def
G	50-90↑*	10-50↓	10-50↑*	0-10*	0-10*	90-100	50-90↑	50-90	Inc.*	0-10*	90-100*	10-50↑	50-90*	0-10*	0-10*
H	0-10*	0-10*	0-10*	0-10*	0-10*	10-50↓*	0-10*	0-10↑*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*
C	90-100*	0-10*	10-50↑*	0-10*	50-90	90-100*	50-90↓	50-90↑*	10-50↑*	50-90↑*	90-100*	10-50↑*	50-90↓*	Inc.*	10-50↑*
F	Inc.*	0-10*	10-50↓*	0-10*	10-50↓*	90-100	50-90	0-10*	0-10*	90-100	Inc.*	Inc.*	Inc.*	0-10*	50-90↓*
A	Inc.*	0-10*	0-10↑*	0-10*	10-50↓*	90-100	Inc.	0-10*	0-10*	90-100	Inc.*	Inc.*	Inc.*	0-10*	Inc.*
B	90-100*	50-90	10-50*	0-10*	0-10*	90-100	50-90	50-90↑	50-90	0-10*	90-100*	50-90*	50-90*	0-10*	0-10*
O	90-100	50-90	Inc.*	0-10*	90-100*	90-100*	Inc.	50-90↓*	0-10*	50-90↓*	0-10*	0-10*	0-10*	0-10*	0-10*
M	0-10*	0-10*	Inc.*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*	0-10*

Detectors are listed row-wise, while image transformations are listed column-wise. Features are scored based on their matching performance (%TP): 0–10% (red), 10–50% (orange), 50–90% (yellow), and 90–100% (green). All detectors are labeled with their initials except DoG(G) and DoH(H). Due to the large detection times, MSER results are computed over a reduced dataset of ten representative images. Table better seen in color.

TABLE IV
VALIDATION OF FOUR SELECTED FEATURES IN THE *KneeValidation dataset*

Feature	Metric	Subsets										Mean
		A	B	C	D	E (M)	F (M)	G (M)	H (M)	I (L)	J (M,L)	
DoG+SIFT	%TP	100%	100%	100%	100%	100%	62%	96.7%	100%	74%	91.7%	**92.4%**
	# of matches	21	258.4	634.8	41.2	13	6.8	47.4	60.6	10.5	10.5	**110.4**
DoG+SURF	%TP	64%	63.3%	100%	40%	30%	20%	74.3%	60%	30.4%	31.3%	**51.3%**
	# of matches	16	133.8	342	18.8	7.8	8	31.2	28.4	5.3	7.3	**59.9**
BRISK+SURF	%TP	100%	100%	100%	100%	100%	51.1%	95.3%	100%	94.5%	87.5%	**92.8%**
	# of matches	22	68.8	61	102	9.4	8.2	25.2	19.6	10	7.5	**33.4**
BRISK+BRISK	%TP	85.9%	100%	100%	100%	100%	40%	96%	100%	88.2%	95.8%	**90.6%**
	# of matches	12.4	70.2	58.2	66.6	5.6	3.4	17	16	8.8	4.5	**26.3**

Subsets include assorted combinations of motion (M) and lighting (L) variations.

source's brightness was varied across its full adjustable range, we did not observe a correlated performance reduction.

In summary DoG+SIFT seems to be the best feature for matching arthroscopic images. This result is consistent with the matching and tracking robustness reported in the literature across multiple applications with variegated datasets. However, while the 128-byte descriptor encodes most of a features' saliency, it could also restrict the usability in real-time applications. Features BRISK+BRISK and BRISK+SURF offer an interesting balance between matching performance and computation time thanks to a fast keypoint detector and compact 64-byte descriptors. For instance, the BRISK+BRISK feature is reported to be up to 100 times faster than its DoG+SIFT counterpart [28].

VI. CONCLUSION

Robotic surgical assistants will revolutionize the way knee arthroscopy is performed. Techniques such as SFM, visual SLAM and visual servoing will require robust, accurate and fast feature matching algorithms. As a first crucial step, this letter presented the first detailed experimental evaluation of the performance of state-of-the-art keypoint detection and description methods on knee arthroscopic images.

Firstly, the influence of parameter settings on *repeatability*, *numKeypoints* and *1-spread* was evaluated for eight state-of-the-art detectors. Secondly, the matching performance of features obtained by combining these eight detectors with six state-of-the-art descriptors was evaluated on a dataset of arthroscopic images undergoing a variety of synthetic transformations. Suitable candidates were further validated under real imaging conditions.

Our results showed that, with a suitable parameter configuration, state-of-the-art features can be used to describe and correctly match arthroscopic images. DoG+SIFT led to the best matching performance, i.e. highest rate of correct correspondences that were also well spread in the image. This feature is best suited for matching-based techniques that are executed offline. BRISK+BRISK and BRISK+SURF exhibited slightly lower performance, however, these methods have the potential of being executed orders of magnitude faster [28]. Consequently, this work recommends the use of the latter two features for viable implementations of real-time matching algorithms for arthroscopy.

In future work, a visual SLAM algorithm will be built exploiting the performance of the recommended features. Initially, we will focus on the exploratory stage of arthroscopy when tools are not present to deform or resect tissues. Similarly, we will assume that the patient's leg is not repositioned during the scene reconstruction. In latter stages of the project we will gradually remove these assumptions.

Other topics to explore include the automatic segmentation of tissue and surgical tools in arthroscopic images, the

investigation of whether learnt features can outperform the ones recommended by this letter, and the extension of the letter's outcomes to other body joints (e.g. hip or shoulder).

REFERENCES

[1] D. Vermesan and R. Prejbeanu, "Operating setup and normal anatomy," in *Atlas of Knee Arthroscopy*. London. U.K.: Springer, 2015, pp. 1–17.

[2] J. W. Griffin, J. A. Hart, S. R. Thompson, and M. D. Miller, "Basics of knee arthroscopy," in *DeLee & Drez's Orthopaedic Sports Medicine*. Amsterdam, The Netherlands: Elsevier, 2014.

[3] L. Maier-Hein *et al.*, "Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery," *Med. Image Anal.*, vol. 17, no. 8, pp. 974–996, 2013.

[4] M. Azizian, M. Khoshnam, N. Najmaei, and R. V. Patel, "Visual servoing in medical robotics: A survey. Part I: endoscopic and direct vision imaging techniques and applications," *Int. J. Med. Robot. Comput. Assisted Surg.*, vol. 10, no. 3, pp. 263–274, 2014.

[5] L. Maier-Hein *et al.*, "Comparative validation of single-shot optical techniques for laparoscopic 3-D surface reconstruction," *IEEE Trans. Med. Imag.*, vol. 33, no. 10, pp. 1913–1930, Oct. 2014.

[6] S. Gauglitz, T. Hllerer, and M. Turk, "Evaluation of interest point detectors and feature descriptors for visual tracking," *Int. J. Comput. Vision*, vol. 94, no. 3, pp. 1913–1930, Oct. 2011.

[7] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[8] O. Grasa, E. Bernal, S. Casado, I. Gil, and J. Montiel, "Visual SLAM for handheld monocular endoscope," *IEEE Trans. Med. Imag.*, vol. 33, no. 1, pp. 135–146, Jan. 2014.

[9] A. Agudo, F. Moreno-Noguer, B. Calvo, and J. Montiel, "Sequential non-rigid structure from motion using physical priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 979–994, May 2016.

[10] N. Mahmoud *et al.*, "ORBSLAM-based endoscope tracking and 3D reconstruction," in *Computer-Assisted and Robotic Endoscopy*. New York, NY, USA: Springer, 2017, pp. 72–83.

[11] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Found. Trends Comput. Graph. Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[12] K. Mikolajczyk *et al.*, "A comparison of affine region detectors," *Int. J. Comput. Vision*, vol. 65, nos. 1/2, pp. 43–72, 2005.

[13] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[14] P. Mountney, B. Lo, S. Thiemjarus, D. Stoyanov, and G. Zhong-Yang, "A probabilistic framework for tracking deformable soft tissue in minimally invasive surgery," *Med. Image Comput. Comput. Assisted Intervention*, vol. 10, pp. 34–41, 2007.

[15] S. Giannarou, M. Visentini-Scarzanella, and G. Yang, "Affine-invariant anisotropic detector for soft tissue tracking in minimally invasive surgery," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2009, pp. 1059–1062.

[16] 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 Trans. Med. Imag.*, vol. 31, no. 11, pp. 2169–2182, Nov. 2012.

[17] X. Du *et al.*, "Robust surface tracking combining features, intensity and illumination compensation," *Int. J. Comput. Assisted Radiol. Surg.*, vol. 10, no. 12, pp. 1915–1926, 2015.

[18] T. Bergen and T. Wittenberg, "Stitching and surface reconstruction from endoscopic image sequences: A review of applications and methods," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 1, pp. 304–321, Jan. 2016.

[19] M. Lourenco, J. P. Barreto, and F. Vasconcelos, "sRD-SIFT: Keypoint detection and matching in images with radial distortion," *IEEE Trans. Robot.*, vol. 28, no. 3, pp. 752–760, Jun. 2012.

[20] P. Abeles, "Boofcv v0.25," Dec. 23, 2016. [Online]. Available: http://boofcv.org/

[21] A. Vedaldi and B. Fulkerson, "VLFeat," Dec. 23, 2016. [Online]. Available: http://www.vlfeat.org/

[22] "*OpenCV v3.1.0*," Nov. 23, 2016. [Online]. Available: http://opencv.org//

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

[24] H. Bay, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vision Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[25] M. Agrawal, K. Konolige, and M. Blas, "Censure: Center surround extremas for realtime feature detection and matching," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 102–115.

[26] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. 9th Eur. Conf. Comput. Vision*, 2006, pp. 430–443.

[27] E. Mair, G. Hager, D. Burschka, M. Suppa, and G. Hirzinger, "Adaptive and generic corner detection based on the accelerated segment test," in *Proc. Eur. Conf. Comput. Vision*, 2010, pp. 183–196.

[28] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 2548–2555.

[29] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vision*, 2011, pp. 2564–2571.

[30] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vision Comput.*, vol. 22, no. 10, pp. 761–767, 2004.

[31] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. Europ. Conf. Comput. Vision*, 2010, pp. 778–792.

[32] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 510–517.

[33] P. Mountney, D. Stoyanov, A. Davison, and G. Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," in *Medical Image Computing and Computer-Assisted Intervention*. New York, NY, USA: Springer, 2006, pp. 347–354.

[34] P. H. Torr, "Bayesian model estimation and selection for epipolar geometry and generic manifold fitting," *Int. J. Comput. Vision*, vol. 50, no. 1, pp. 35–61, 2002.