The International Journal of **Medical Robotics** and **Computer Assisted Surgery**

THE INTERNATIONAL JOURNAL OF MEDICAL ROBOTICS AND COMPUTER ASSISTED SURGERY
Int J Med Robotics Comput Assist Surg 2015.
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/rcs.1661

REVIEW ARTICLE

Video-based 3D reconstruction, laparoscope localization and deformation recovery for abdominal minimally invasive surgery: a survey

Bingxiong Lin[1]
Yu Sun[1]*
Xiaoning Qian[2]
Dmitry Goldgof[1]
Richard Gitlin[3]
Yuncheng You[4]

[1] *Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA*

[2] *Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA*

[3] *Department of Electrical Engineering, University of South Florida, Tampa, FL, USA*

[4] *Department of Mathematics and Statistics, University of South Florida, Tampa, FL, USA*

*Correspondence to: Yu Sun, 4202 E. Fowler Avenue, ENB 118, Tampa, FL 33620, USA.
E-mail: yusun@cse.usf.edu

Accepted: 23 March 2015

Abstract

Background The intra-operative three-dimensional (3D) structure of tissue organs and laparoscope motion are the basis for many tasks in computer-assisted surgery (CAS), such as safe surgical navigation and registration of pre-operative and intra-operative data for soft tissues.

Methods This article provides a literature review on laparoscopic video-based intra-operative techniques of 3D surface reconstruction, laparoscope localization and tissue deformation recovery for abdominal minimally invasive surgery (MIS).

Results This article introduces a classification scheme based on the motions of a laparoscope and the motions of tissues. In each category, comprehensive discussion is provided on the evolution of both classic and state-of-the-art methods.

Conclusions Video-based approaches have many advantages, such as providing intra-operative information without introducing extra hardware to the current surgical platform. However, an extensive discussion on this important topic is still lacking. This survey paper is therefore beneficial for researchers in this field. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords visual SLAM; surface reconstruction; tissue deformation; feature detection; feature tracking; laparoscopy; pose estimation; camera tracking

Introduction

Compared with traditional open-cavity surgeries, the absence of large incisions in minimally invasive surgery (MIS) benefits patients through smaller trauma, shorter hospitalization, less pain and lower infection risk. In abdominal MIS, the abdomen is insufflated and surgeons gain access to the tissue organs through small incisions. Laparoscopic videos provide surgeons with real-time images of surgical scenes, based on surgical instruments that are precisely manipulated. However, laparoscopic videos are two-dimensional (2D) in nature, which poses several restrictions on surgeons. For example, depth information is lost in 2D images, so surgeons have to estimate the depth, based on their experience. Stereo images from stereoscopic laparoscopes must be fed to the left and right eyes separately to have a sense of depth, whereas depth information exists only in the surgeon's mind and has not yet been explicitly calculated (1). In addition, a laparoscope gives a narrow field of view, which makes it difficult for surgeons to understand the position and orientation of the laparoscope and the surgical instruments (2). Moreover, currently, the most significant limitation of an endoscopic video is that it is unable to provide any information about tissue structures underneath organ surfaces. For example, in colon surgeries, surgeons have to spend a large amount of time dissecting tissues to identify the ureters underneath.

To overcome the inherent restriction of 2D laparoscopic videos, computer-assisted surgery (CAS) has been proposed to guide a surgical procedure by providing accurate anatomical information about the patient. In CAS, a key step is to register pre-operative data, such as magnetic resonance imaging (MRI) and computed tomography (CT), with intra-operative data, so that the pre-obtained patient anatomy can be accurately displayed during surgery. Registration has been a longstanding research topic in the literature. Notably, registration in neurosurgery has become successful due to the availability of fixed structures, such as bones. A wide range of medical image registration methods and augmented reality techniques have been proposed for neurosurgery. Maintz and Viergever (3) presented a comprehensive survey on medical image registration methods and provided nine criteria to classify those methods into different categories. One major dichotomy used in (3) was whether those obtained correspondences were from extrinsic or intrinsic sources; extrinsic registration methods rely on foreign objects, such as fiducial markers, and intrinsic methods are based on anatomical structures. Another survey on medical image registration is available in (4). Recently, Markelj *et al.* (5) provided a detailed review on registration of three-dimensional (3D) pre-operative data and 2D intra-operative X-ray images.

A fundamental task in medical image registration is to overlay images of the same scene taken at different times or from different modalities. Many methods have been proposed, and a survey paper (6) for neurosurgery has been presented. Another important task of medical image registration is to overcome the morphology issues of soft tissues, such as the brain and the lung, which might shift and deform and cause error to the global rigid registration. In the research community, multiple survey papers (7,8) on deformable medical image registration have been presented. Thorough evaluation experiments of localization and registration accuracy in clinical neurosurgery are available (9,10).

Despite the success of medical image registration in neurosurgery, its application in abdominal surgery has presented many challenges, due to the deforming environment of the abdomen. It is difficult to find rigid anatomical landmarks on the abdomen because the abdominal shape changes after gas insufflation. Moreover, even if the global patient–CT registration is available, registration in the abdominal area is not likely to be accurate because tissues and organs can easily slide and deform, due to gas insufflation, breathing or heartbeats. To overcome these limitations and the challenges of registration in the abdominal environment, intra-operative 3D reconstruction of surgical scenes and laparoscope localization, based on video content, are the fundamental tasks in CAS for abdominal MIS. For example, recovery of the time-varying shapes of the deforming organs can be used to determine the tissue morphology. Laparoscope localization can help surgeons determine where the instruments are when operating, with respect to the human anatomy.

Different methods of vision-based 3D reconstruction and laparoscope localization have been proposed in the literature. However, large-area 3D reconstruction, laparoscope localization and tissue deformation recovery in the abdominal environment in real time remain open challenges to researchers. The difficulties are mainly from the special environment of the abdominal MIS. First, compared with general images taken in a man-made environment, MIS images usually contain homogeneous areas and specular reflections, due to the smooth and wet tissue surface (11,12). These properties significantly affect the performance of state-of-the-art feature point detection methods. Without reliable feature-point correspondences, many feature-based 3D reconstruction and visual simultaneous localization and mapping (SLAM) (13–15) methods developed in computer vision do not perform well. Second, surgical scenes are highly dynamic and change from time to time during a surgical procedure, e.g. surgical instruments are moving in the surgical site and may cause occlusion problems, and soft tissues may have non-rigid deformation due to respiration or interaction with surgical instruments. In the dynamic and non-rigid MIS environment, simultaneous 3D reconstruction, laparoscope localization and deformation recovery in real time are very difficult (16,17). This problem is referred to as minimally invasive surgery visual SLAM (MIS–VSLAM) in this paper. The purpose of this review is to provide a comprehensive survey of the current state-of-the-art MIS–VSLAM methods for the abdominal MIS environment.

Focus and outline

The remainder of this article is organized as follows. First, as fundamental tasks in 3D reconstruction and laparoscope localization, feature detection and feature tracking methods are discussed. The discussion is focused on how these detection and tracking methods are designed to overcome the difficulties of MIS images, such as low contrast, specular reflection and smoke. Next, laparoscopic

DOI: 10.1002/rcs

video-based 3D surgical scene reconstruction methods without the estimation of camera motion are introduced, and are summarized based on the adopted vision cues, such as stereo, structured light and shadow. Note that, in addition to the challenges from feature detection, 3D reconstruction methods in MIS must overcome extra difficulties from surgical instrument occlusion, the small baseline of stereo cameras and the constrained environment. Then, the camera motion is estimated during the 3D reconstruction, and the scene is assumed to be rigid or static. With the rigid scene assumption, visual SLAM becomes relatively easier, and many methods have been presented in the computer vision and robotics literature. These methods and how they are applied in MIS to overcome the corresponding difficulties are discussed. Finally, the most difficult problem is considered – visual SLAM in dynamic and deforming surgical scenes. This research problem is similar to non-rigid structure from motion (NRSFM) in computer vision. Various approaches have been presented to tackle the problem from different perspectives; these methods are summarized and their key ideas are explained. The classification of MIS–VSLAM methods based on camera motion and scene type is shown in Figure 1.

Materials and methods

This section focuses on the introduction of state-of-the-art methods in image feature detection and tracking, 3D reconstruction, deformation recovery and visual SLAM for abdominal MIS. The review follows the organization shown in Figure 1.

Abdominal MIS set-up and datasets

Typical abdominal MIS set-up

In MIS, multiple ports are needed for the insertion of the laparoscope and surgical instruments. The laparoscope usually has an attached monocular camera. Different monocular laparoscopes might have different angles and light configurations. Stereoscopic laparoscopes are widely used in robotic surgery platforms, such as the da Vinci surgical system (18). The intrinsic and extrinsic parameters of the cameras attached at the tip of laparoscopes can be calculated following the calibration procedure in (19). A diagram of the typical MIS set-up is shown in Figure 2.

Public MIS datasets

Public MIS datasets are valuable to the research community, and multiple MIS datasets have been collected and made available. Hamlyn Centre laparoscopic/endoscopic video datasets (20) contain a large collection of MIS videos for different organs, including lung, heart, colon, liver, spleen and bowel; the videos in (20) include a variety of endoscope motions and tissue motions. Bartoli (21) provided a uterus dataset, which contains tissue deformation caused by instrument interactions. In (12,22) an image dataset was collected for evaluation of the repeatability of feature detectors. The dataset in (22) contains hundreds of images sampled from *in vivo* videos

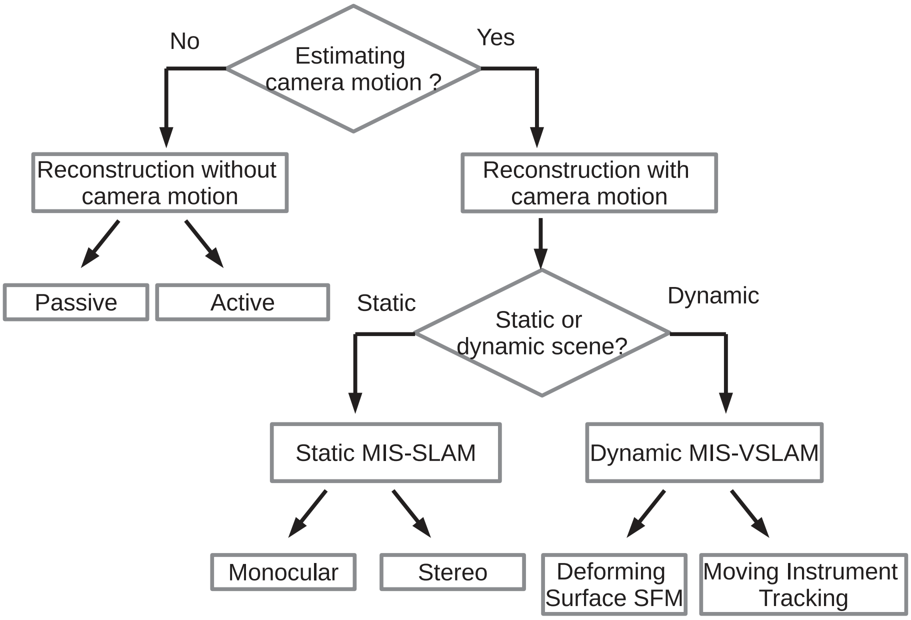

Figure 1. Classification of MIS–VSLAM methods, based on camera motions and scene types

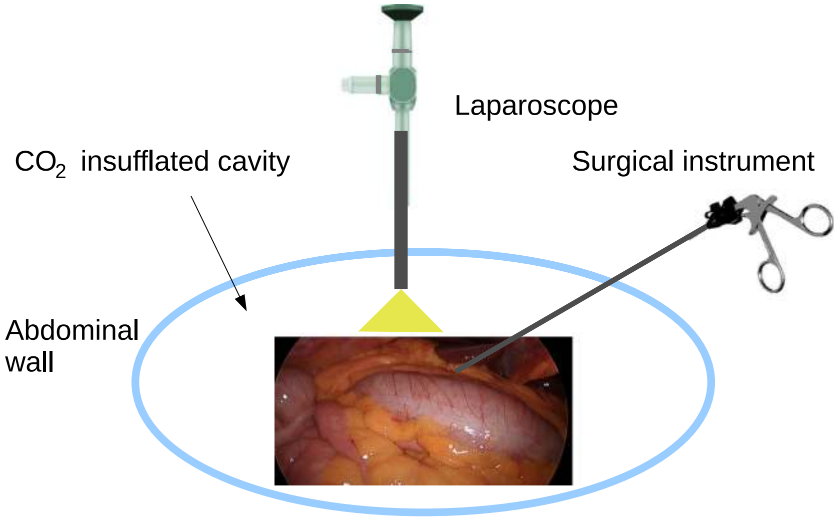

Figure 2. Typical abdominal MIS set-up

taken during colon surgeries; the images in this dataset were taken at different viewpoints, and the ground truth homography mappings are available. Puerto-Souza and Mariottini (23) provided the hierarchical multi-affine (HMA) feature matching toolbox for MIS images, which contains 100 image pairs representing various surgical scenes, such as instrument occlusion, fast camera motion and organ deformation. Stereo videos with surgical instruments moving in front of the liver were made publicly available with ground-truth information of the pose and position of those instruments (24). The Johns Hopkins University Intuitive Surgical Inc. Gesture and Skill Assessment Working Set (JIGSAWS) (25) contained stereo-videos of three elementary surgical tasks on a bench-top model: suturing, knot-tying and needle-passing. The goal of the JIGSAWS dataset was to study and analyse surgical gestures. The Open-CAS (26,27) collected multiple datasets for validating and benchmarking CAS, including liver simulation, liver registration and liver 3D reconstruction. There are also multiple retinal datasets that are publicly available, including structured analysis of the retina (STARE), digital retinal images for vessel extraction (DRIVE) and retinal vessel image set for estimation of widths (REVIEW). A summary of these datasets is shown in Table 1.

Feature detection and feature tracking

Image feature detection and feature tracking are fundamental steps in many applications, such as structure and pose estimation, deformation recovery and augmented reality. Many well-known feature detectors and feature descriptors have been presented. In this section, different feature detection and feature-tracking methods are introduced, and how they are adapted for MIS images is discussed.

Feature detection

Depending on what information is used, feature detection methods can be broadly classified into three categories: intensity-based detectors, first-derivative-based detectors and second-derivative-based detectors. In the first category, feature detectors are mostly based on pixel intensity

Table 1. Summary of publicly available MIS datasets

Datasets	Sensor	Video/image	Scene	Scene motion	Resolution
Hamlyn (20)	Mono and stereo	Video	Abdomen	Rigid and deforming	Varied
Bartoli (21)	Mono	Video	Uterus	Deforming	1280 × 720
Lin *et al*. (12,22)	Mono and stereo	Image	Abdomen	Rigid	Varied
HMA (23)	Mono	Image	Abdomen	Rigid and deforming	704 × 480
Allan *et al*. (24,28)	Stereo	Video	Abdomen	Rigid	1920 × 1080
JIGSAWS (25)	Stereo	Video	Lab	Deforming	640 × 480
Open-CAS (26,27)	Stereo	Image	Liver	Rigid	720 × 576
STARE (29)	Mono	Image	Retina	Rigid	700 × 605
DRIVE (30)	Mono	Image	Retina	Rigid	565 × 584
REVIEW (31)	Mono	Image	Retina	Rigid	Varied

comparisons. In the features from accelerated segment test (FAST) (32), Rosten *et al*. replaced the disk with a circle and detected corner points by identifying the pattern of a continuously bright or dark segment along the circle. Different from FAST, Mair *et al*. introduced a new circle pattern and used a binary decision tree for the corner classification (33).

In the second category, the first derivatives along the x and y coordinates in the raw image, namely I_x, I_y, reflect the intensity change and can be used to detect object structures, such as edges and boundaries. Most methods in this category are based on the eigenvalues of the auto-correlation matrix (34). Harris and Stephens (34) proposed a measure based on those eigenvalues to detect image patches that are likely to be corners. Shi and Tomasi (35) argued that λ_1 itself was a good indicator for corners. Mikolajczyk and Schmid (36) extended the Harris corner detector in scale space and proposed the Harris-affine detector, which had better invariance property under affine transformation. The anisotropic feature detector (AFD) exploited the anisotropism and gradient information to detect interest points (37,38).

In the third category, feature detectors exploit the second derivatives of raw images to detect interest points defined by blobs and ridges. Most methods in this category are based on analysis of the Hessian matrix (22). In the Hessian affine detector (39), the determinants of the Hessian matrices were calculated for all pixels, and the local maxima were selected as feature points. Lowe approximated the Laplacian of Gaussian with the difference of Gaussian (DoG) (40) and built a pyramid image space to detect interest points. The speeded-up robust features (SURF) feature detector replaced the Gaussian filters with box filters to obtain a faster speed (41). It has been reported that general feature point detectors do not perform well in MIS images (22,37). Lin *et al*. (22) observed that there were abundant blood vessels in MIS images and proposed to explicitly detect vessel features. Two vessel features were proposed, namely branching points and branching segments, and thorough experiments verified that the vessel features are more robust and distinctive than general features in MIS images (22). Example of branching points, branching segments and half-branching segments are shown in Figure 3.

It is well known that the performance of feature detectors is determined by multiple parameters, such as the standard deviation of Gaussian smoothing, the discrete quantization of orientation, and the number of bins in the histogram of orientation. Most of the above-mentioned feature detection methods require manual parameter tuning based on personal experience. Stavens and Thrun (42) proposed an unsupervised method that learned those parameters from video sequences. In (42), Harris corners (34,35) were detected and tracked by Lucas–Kanade (LK) optical flow (43), and the patches were stored as training data. The idea of treating feature matching as a classification problem was first introduced by Lepetit *et al*. (44). The synthesized images were used to generate local feature point patches as training data for classification. In early work (45,46), randomized trees were used as the classifier. Later, it was shown that the good performance of feature matching was mainly from the randomized binary tests, rather than the randomized tree classifier and, hence, simple semi-naive Bayesian classifier was adopted (47,48).

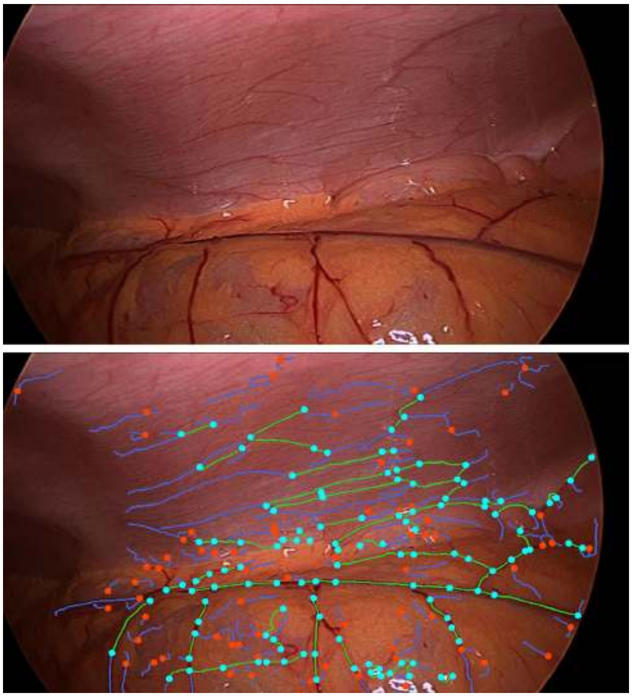

Figure 3. Branching points (cyan), branching segments (green) and half-branching segments (blue) (22): (top) original image; (bottom) image with detected branching segments

Feature tracking

To track feature points, the target feature points are usually represented by their local image patches. Based on local patch representations, tracking methods can be broadly classified into two categories: intensity-based tracking and descriptor-based tracking. In the first category, each feature point is directly represented by the intensity values of the pixels in its local square patch. By assuming that each pixel has a constant intensity, the well-known LK tracking method (43) compares and matches image patches in successive frames, using the sum of squared difference (SSD). To incorporate temporal information, many methods exploit the motion constraints and estimate the probabilities of matches, such as in extended Kalman filter (EKF) (15). In the MIS environment, the tissues might have deformation and the surgical instruments might cause occlusion problems. Mountney and Yang (49) proposed an on-line learning mechanism and treated the tracking as a classification problem. The thin plate spline (TPS) model was successfully applied (50) to track a region of a deforming surface. Richa *et al*. (51) extended the work in (50) to track the heart surface with stereo cameras. Many other tracking methods were introduced and compared in (38).

The recovery of heart motion is a fundamental task in cardiac surgery, and feature tracking using stereo images from stereoscopic laparoscopes has shown promising results. Note that there are two kinds of feature matching with stereoscopic laparoscopes: temporal matching, for successive frames, and spatial matching between left and right images. Typically, feature points are detected in both left and right images, and feature points in the first frame are matched temporally with successive frames to enable tracking. Stoyanov *et al*. (52) used a Shi–Tomasi detector (35) and an MSER descriptor (53) to perform spatial matching. The LK tracking (43) framework was used to track the initial features, and the intensity information of both stereo images was used during the estimation of the warp (52). It was reported in (54) that the use of the LK tracking framework was not very stable, due to the large tissue motion and the fact that some feature points were not well tracked. In (55), feature-based methods (52) and scale-invariant feature transform (SIFT) (40) are combined with intensity-based methods (51) to generate a hybrid tracker for the purpose of robustness.

Since pixel intensities used in the first category are sensitive to lighting conditions, most of these methods make it difficult to track features across large viewpoint changes. On the other hand, in the second category, feature-tracking methods are reliant on feature descriptors to represent feature points. Many feature descriptors have been presented, such as SURF (41), SIFT (40) and binary robust independent elementary features (BRIEF) (56). Feature descriptors are usually normalized and processed to overcome problems such as illumination and appearance changes. As a result, they are usually more robust than the intensity comparison in LK-based tracking. Due to the special environment of MIS, descriptor-based feature matching is not robust towards large viewpoint changes. To overcome this problem, different methods have been introduced to exploit the geometrical properties of the tissue surface. Puerto Souza *et al*. (23,57,58) clustered feature points into different groups, and the local area of each cluster was assumed to be planar. Lin *et al*. (59) first obtained a 3D tissue shape using the TPS model on stereo images and then used the estimated 3D shape to improve feature point matching over large viewpoint changes. A comprehensive study on the evaluation of different feature descriptors on MIS images was reported in (60).

One major challenge of descriptor-based tracking is the time-consuming calculation and matching of descriptors. Currently, without special hardware such as graphics processor units (GPUs), the SIFT feature extraction is still difficult for achieving real-time speed. Recently, from the speed point of view, Yip *et al*. (61,62) proposed a significant tracking-by-detection method that achieved a speed of 15–20 Hz on a MIS scene with tissue deformation and instrument interaction. The major novelty of the method presented in (61,62) is that a feature list is dynamically maintained and updated, which makes it robust to large deformation and occlusion. In (61), for speed consideration, the Star detector (63) implementation of the centre surround extremas (CensurE) (64) feature detector and BRIEF descriptor were used. To further speed up the tracking process, prior information of the surgical scene, such as small camera motion and small-scale change, was exploited to reduce the unnecessary feature comparisons (61). An extensive comparison of tracking accuracy and speed among Star + BRIEF, SIFT and SURF was provided in (62).

To evaluate feature detection and feature tracking, one key task is to generate ground-truth point correspondences across multiple views: to obtain these, typically, experienced human subjects are trained to select the same scene point in multiple images. However, the ground-truth information sometimes might not be sufficiently accurate, due to the manual selection process. To minimize ground-truth error, Maier-Hein *et al*. (65) extended a crowd sourcing-based method to generate reference correspondences for endoscopic images. The correspondence error was reduced from 2 pixels to 1 pixel after applying crowd sourcing (66).

After the ground-truth point correspondences are generated for each frame, the evaluation of feature point tracking can be successfully carried out (38). To evaluate the feature point detection, traditional methods such as (39) usually rely on planar scenes, so that global homography mappings are available. In (22), homography mappings were obtained for flat tissue surfaces, such as the abdominal wall, to evaluate the repeatability of bifurcations (branching points). However, the scenes were not strictly planar, and the homography mappings were not accurate enough to evaluate general feature points that were smaller than branching points (22). Klippenstein and Zhang (67) estimated the fundamental matrices between the first frame and other frames and defined the distances of feature points to the epipolar lines as the error for feature tracking. Different feature detectors and feature-matching methods have been compared in (67); however, the mappings used in (67) are not bijective and, therefore, the definition of error is not accurate. Selka *et al*. (68) reported a forward–backward tracking method for evaluation of both feature detectors and feature tracking. In (68), the MIS video sequence was reorganized into the order: $(I_0, I_2, \ldots, I_{n-2}, n, I_{n-1}, .., I_3, I_1, I_0)$. Those points that were detected in both the first and last frames were called robust points, and the percentage of robust points was used to represent the performance of feature detector and feature tracking.

Discussion

Feature detection and feature tracking are well-studied topics in computer vision. However, distinctive feature detection, matching and tracking for endoscopic images are still challenging, due to the special features of the endoscopic environment, such as poor texture, bleeding, smoke and moving light sources. One future research direction is to exploit the special structures shown in laparoscopic images, such as blood vessels and blood dots caused by surgical instruments. An image feature detector tuned specifically for blood vessels (22) has shown promising results. Since light sources are mounted at the tip of a laparoscope, the light illumination is non-uniform and increases the difficulty in finding the image-point correspondences. As pointed out in (22), laparoscopic images usually have stronger lighting in the centre than at the borders. It is interesting to look into how to remove or reduce the influence from this non-homogeneous illumination from the endoscopic lighting. Another promising research direction is to integrate supervised learning techniques into feature detection and tracking, such as the work in (49).

3D reconstruction without camera motion

In this section, 3D surface reconstruction methods without the consideration of camera motion are introduced. These methods are separated into different categories, based on the vision cues applied.

Stereo cue

Stereo laparoscopes have become widely used in robotic surgery platforms, such as the da Vinci system, to provide 3D views for surgeons. Since no extra hardware is required, reconstruction using stereo laparoscopes has been considered to be one of the most practical approaches for MIS (18). Lau *et al.* (69) used the zero mean sum of squared difference (ZSSD) for stereo matching and, later, the heart surface was estimated using the B spline-based method (69). Kowalczuk *et al.* (70) evaluated the stereo-reconstruction results of the operating field with porcine experiments.

Recently, Stoyanov *et al.* (18) proposed a novel stereo matching algorithm for MIS images, which was robust to specular reflections and surgical instrument occlusion. They proposed to first establish a sparse set of correspondences of salient features and then propagate the disparity information of those salient features to nearby pixels. The propagation in (18) was based on the assumption that the disparity values of the nearby pixels in MIS images were usually very similar, since many tissue organ surfaces are locally smooth. Stereo-reconstruction of the liver surface is known to be difficult because of the homogeneous texture. Totz *et al.* (71) proposed a semi-dense stereo-reconstruction method for liver surface reconstruction, which adopted a coarse-to-fine pyramidal approach and relied on GPU to exploit the parallelism. In (72), semi-dense stereo-reconstruction results (18) from different viewpoints were merged to obtain large-area 3D reconstruction results of the surgical scene, based on camera localization results from (16). In (73), the local surface orientation was estimated based on the constraints from the endoscope camera and light sources, and then fused with the semi-dense reconstruction from (18) to generate a gaze-contingent dense reconstruction. Stoyanov (17) reported a 3D scene flow method to estimate the structure and deformation of the surgical scene by imposing spatial and temporal constraints.

Distinctive feature points can be matched in stereo images to obtain a set of sparse 3D points. To achieve dense reconstruction results of tissue surfaces, different methods have been proposed to incorporate geometrical constraints of tissue surfaces. Richa *et al.* (55) tracked feature points over stereo images and obtained the 3D positions of those feature points based on triangulation. Later, the sparse 3D points were chosen as the control points in a TPS model, and a dense 3D shape was estimated (55). Bernhardt *et al.* (74) analysed the surgical scenes and presented three criteria for stereo matching to remove outliers. After the outliers were discarded, the holes were filled with the median of their neighbouring pixel values (74). Chang *et al.* (75) first obtained a coarse reconstruction using the zero-mean normalized cross-correlation (ZNCC) and then refined the disparity function using a $Huber - L^1$ variational functional.

Active methods

Most of the above methods are dependent on the texture of tissue surfaces to establish feature-point correspondences for reconstruction. These methods become unstable if tissue surfaces are poorly textured. To overcome this problem, many methods aim to actively project special patterns, using laser stripes or structured light, onto tissue surfaces and build correspondences based on those patterns. When stereo cameras are available, the light source does not need to be calibrated and, therefore, the system becomes relatively easy to use (11). Otherwise, the Euclidean transformation between the monocular camera and the light source needs to be accurately calibrated, and after calibration the system has to be fixed during the whole reconstruction procedure.

Different methods have been proposed to project laser stripes on organ surfaces for reconstruction. In (76), a laser stripe was projected in the laparoscopic environment to measure intracorporeal targets. To measure the 3D shape of the surgical site in real time, a laser-scan endoscope system with two ports was designed (77). For the calibration of this system, infrared markers were placed at the ends of both the camera device and the laser device and tracked using the OPTOTRAK system (77). The root mean square error of measurements among those markers was reported to be 0.1 mm (77).

Instead of using laser stripes, other methods project an encoded light pattern on tissue surfaces. Different light patterns have been designed to establish the correspondences between the camera and the projector (78,79). To recover the dynamic internal structure of the abdomen in real time, Albitar *et al*. (80) developed a new monochromatic pattern composed of three primitives: disc, circle and strip. The images were processed to detect and discriminate the primitives, whose spatial neighbourhood information was used to establish correspondences between the captured image and the known pattern (80). The developed system was able to project 29×27 primitives on an area of size $10 \times 10\,\mathrm{cm}^2$ (80). Later, Maurice *et al*. designed a new spatial neighbourhood-based framework to generate coded patterns with 200×200 features, using the mean Hamming distance (81).

One major challenge of using either laser or structured light is that the whole 3D scanning system is usually too large to fit into the current MIS set-up (82). To overcome this size problem, Schmalz *et al*. (82) designed a very tiny endoscopic 3D scanning system composed of a catadioptric camera and a sliding projector (82). The sensor head in the scanning system had a diameter of 3.6 mm and a length of 14 mm (82). The system was specifically designed for a tubular environment and was able to obtain the 3D depth at 30 fps, with a working cylindrical volume of about 30 mm in length by 30 mm in diameter (82). Clancy *et al*. (83) designed another tiny structured lighting probe with a 1.7 mm diameter; in their system, a set of points were projected and each point was assigned a unique wavelength.

Recently, the time-of-flight (TOF) camera sensor has become popular for 3D reconstruction. Penne *et al*. (84) designed an endoscope system with a TOF camera sensor. Haase *et al*. (85) proposed a method to fuse structures recovered from different frames of a TOF sensor to obtain large-area reconstruction results. More details regarding the TOF-camera-based reconstruction methods can be found in (86).

Shading and shadow cue

As one of the well-studied 3D reconstruction methods in computer vision, shape-from-shading (SFS) is very appealing to researchers because it does not require extra hardware in MIS. Many researchers have attempted to apply SFS to recover the shape from a monocular camera (87). Wu *et al*. first extended the SFS problem to a perspective camera and near-point light sources and then applied it to reconstruct the shape of bones from near-lighting endoscopic video (88). The application of SFS in MIS is difficult and has multiple restrictions. To begin with, endoscopic images do not satisfy the common assumptions required by SFS, Lambertian reflectance and uniform albedo (17). Additionally, with SFS it is generally not possible to recover a complete 3D surface with one lighting condition because each pixel has only one intensity measurement, which is not enough to recover the surface orientation that has two degrees of freedom (89). Therefore, multiple lighting conditions with a constant viewing direction are required to theoretically achieve a complete surface recovery, which is commonly known as ‘photometric stereo’ (PS) (89); please refer to (89,90) for more details about PS.

During the MIS procedure, shadows cast by surgical instruments are good sources of visual cues for reconstruction. Researchers are also interested in generating optimal shadows for MIS surgeries in terms of contrast and location of shadow-casting illumination (91). In (92), an ‘invisible shadow’ was generated by a secondary light source and was detected and enhanced to provide a depth cue. Rather than estimating the position of the light source as in the classic methods in (93), Lin *et al*. (11) proposed to use stereo cameras and mount a single-point light source on the ceiling of the abdominal wall to generate shadows. The borders of the generated shadows were later detected in both stereo images and a dense disparity map was interpolated (11). The shadow-casting process in (11) is illustrated in Figure 4; typical examples of reconstruction results using the Lin *et al*. method are shown in Figure 5. The benefit of using stereo cameras is that the light source is no longer required to be stationary. However, to generate shadows cast by surgical instruments, an extra overhead light source is needed.

Discussion

In stereo reconstruction, because of the similarity of left and right images from stereo cameras, feature point matching between the two channels is relatively easy and a sufficient number of feature point correspondences can be established if rich texture is available. Currently, one of the main challenges in stereo-reconstruction for MIS is how to obtain dense reconstruction results. Interesting future research directions include building suitable models for tissue surfaces and integrating laparoscope

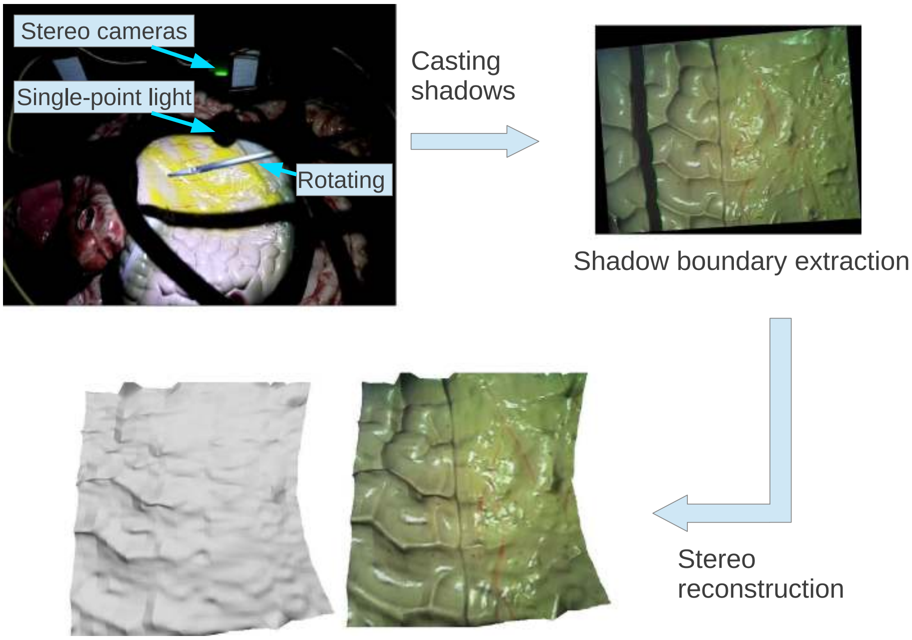

Figure 4. 3D reconstruction method by casting shadows, using surgical instruments from (11)

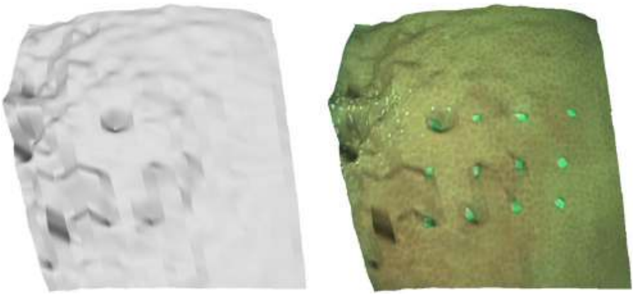

Figure 5. One example of 3D reconstruction results of *ex vivo* porcine liver, with and without texture mapping using shadows and stereo cameras (11)

motions, such as the work in (94). Active methods are able to obtain accurate 3D information without depending on tissue texture and, therefore, are attractive to researchers. The main drawback of the active methods is the requirement of extra hardware in current surgical platforms. In the future, it will be necessary to design very small-scale hardware that is compatible with the MIS surgical platform. Meanwhile, how to generate optimal structure patterns for MIS is also an important research topic (81). Methods based on defocus have also shown the ability to recover the 3D structure of tissue surfaces (95) and need further investigation. To better apply SFS in MIS, a more advanced reflectance model for the laparoscopic environment is needed (86).

Rigid MIS–VSLAM

In the previous section, no camera motion was considered during the 3D reconstruction process, and hence the motion could not be recovered or used. In practice, the endoscopic cameras are usually moving during the MIS procedure and the motion can be used to recover the 3D structure. Additionally, knowledge of the camera pose is crucial to help surgeons better understand the surgical environment. For example, accurate camera tracking is necessary for safe navigation and instrument control during endoscopic endonasal skull base surgery (ESBS) (96).

Many external endoscope tracking methods that rely on passive optical markers have been presented, and have been used to track the location of an endoscope relative to CT. Shahidi *et al.* (97) reported millimetre tracking accuracies of a marker-based external tracking system. Lapeer *et al.* (98) showed that sub-millimetre accuracy was still difficult to achieve. Mirota *et al.* (96) presented an endoscope-tracking method that relies on the video content only and achieved accuracy at 1 mm. Compared with external marker-based tracking systems, video-based endoscope localization has the advantage that no marker or external system is needed. In addition, passive markers might be blocked from the tracking system during surgery and cause tracking failures. Therefore, the combination of external tracking and video-based tracking can potentially offer more robust tracking results. An extensive discussion on external tracking is beyond the scope of this paper;

more detail regarding external tracking is available in (98). From here forward, we focus on video-based camera tracking. In this section, the surgical scene is assumed to be rigid (static) and methods that simultaneously estimate the 3D structure and the camera motion are introduced. The illustration of MIS–VSLAM methods in rigid scenes is shown in Figure 6.

In computer vision, many methods in structure from motion (SFM) have been proposed to estimate the sparse 3D structure of a rigid scene from a set of images taken at different locations (99,100). The technique has been scaled up successfully to a large dataset with millions of images taken from the internet (100). SFM has also been applied in MIS to expand the field of view for surgeons and recover a wide area of 3D structures (101,102). It is known that SFM greatly depends on robust wide-baseline feature matching, such as SIFT on images of man-made buildings. However, wide-baseline feature matching is difficult on low-texture MIS images. Hu *et al*. (54) presented a method to alleviate this problem for totally endoscopic coronary artery bypass (TECAB) surgery (54). A genetic/evolutionary algorithm was proposed (54,103) to overcome the missing data problem during LK tracking. Another drawback of SFM is that it processes all images together to optimize the 3D structures and the cameras' poses. One benefit of this global batch optimization is that the recovered structure and camera poses can achieve high accuracy. However, the number of parameters is large and the optimization requires expensive computation, which makes the system impractical for real-time purposes. To reduce these difficulties, the laparoscope was tracked externally to provide camera poses in the optimization of SFM (104).

Different from SFM, in robotics one main task is to achieve real-time camera localization. Robotics researchers treat the camera as a sensor observing and moving in an explored or unexplored environment, and the problem is normally termed 'visual SLAM'. SLAM is a well-studied topic in robotics and has been applied to the automatic navigation of mobile robots in an unexplored environment. Comprehensive surveys of SLAM can be found in (13,14). Originally, SLAM was designed for range sensors, such as laser range finder and sonar systems, which obtain 3D information with uncertainty directly from the sensor reading. Different from that, a monocular camera is a bearing-only sensor that needs at least two measurements from different locations to calculate the 3D information. However, the availability of cameras and rich information in each image has made the camera a popular sensor for SLAM.

Monocular camera

Burschka *et al*. (105,106) proposed an early framework to simultaneously estimate 3D structures and camera poses, based on the endoscopic video. However, the estimation of camera poses in (105,106) was performed frame-by-frame, using the correspondences detected in successive frames, which might lead to a significant accumulated error. To overcome the aforementioned difficulty of feature matching in MIS images, Wang *et al*. (107) first applied Singular Value Decomposition (SVD) matching (108) on SIFT points to obtain more but less accurate

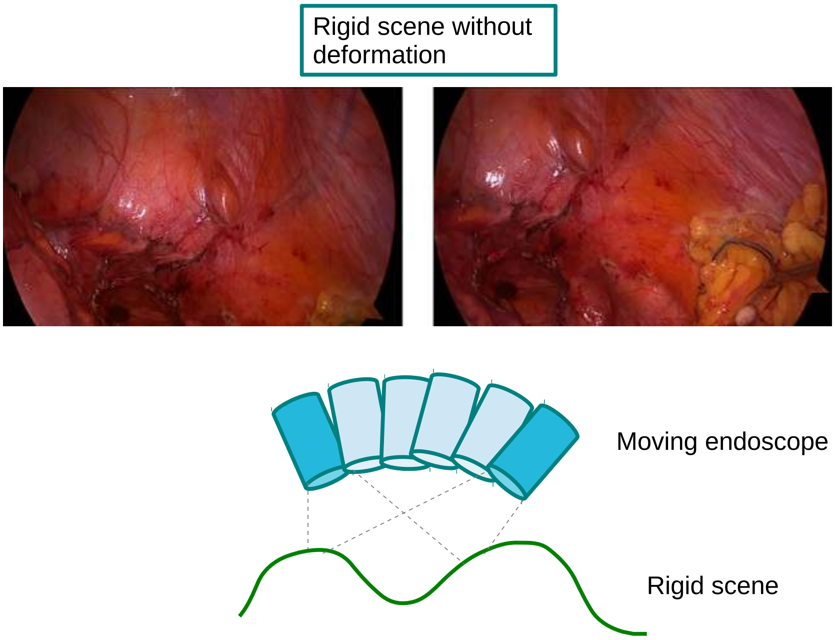

Figure 6. MIS–VSLAM methods in rigid and static scenes; solid green curve represents a rigid and static scene

correspondences, which were further refined by a novel method called 'adaptive scale kernel consensus' (ASKC) (107). With the feature correspondences from successive frames, the method in (107) maintained a 3D feature point list and tracked the camera at each frame. Mori *et al*. (109) designed a visual SLAM system specifically for a bronchoscope (109), in which the motion of the bronchoscope was initially estimated based on the optical flow and was later refined by intensity-based image registration.

The seminal work of Davison (15) was the first significant real-time system that successfully applied the extended Kalman filter–SLAM (EKF–SLAM) framework for a hand-held monocular camera. In (15), feature points were detected by the Shi–Tomasi operator (35) and represented as 2D square patches. The measurement model for the monocular camera first initialized a 3D line when a new feature point was observed, and then calculated the 3D position of the feature point when it was observed the next time (15). Since Davison's system updates the pose and the map at each frame, it can only maintain a small number (typically < 100) of landmarks.

Multiple methods have been introduced to improve Davison's monocular camera EKF–SLAM framework. To overcome the problem of delayed initialization of feature points in (15,110), Civera *et al*. (111,112) presented an inverse-depth parameterization method to unify the initialization and tracking of both close and distant points. Civera *et al*. (113,114) further integrated the random sample consensus (RANSAC) method into the EKF–SLAM framework (15,110) to estimate inliers of feature point matches, and presented the one-point RANSAC method. With the prior information of camera poses, only one sample was needed to initialize model estimation in the RANSAC process, and therefore the RANSAC computation could be greatly reduced (113,114). Based on inverse-depth parameterization (111,112), Grasa *et al*. (115,116) successfully combined the one-point RANSAC method (113) and randomized list relocalization (117) together, so that the system was robust to the challenges from the MIS environment, such as sudden camera motion and surgical instrument occlusion. In a more extensive evaluation of the system from (116), > 15 human ventral hernia repair surgeries were reported in (118), in which the scale information was obtained from the clinch of the surgical instrument. The measurements of the main hernia axes were chosen to represent the accuracy of the reconstruction and the ground truth was measured by tape (118).

In SFM, the time-consuming bundle adjustment has been shown to be very effective in simultaneously optimizing 3D structure and camera poses. To apply bundle adjustment in a real-time system, different methods have been reported and discussed to reduce the computational burden of the bundle adjustment in robotics. Local bundle adjustment was used in (119) to achieve accurate reconstruction results and simultaneously reduce the computation. Later, Klein and Murray (120) introduced the breakthrough work, parallel tracking and mapping (PTAM), which was able to robustly localize the camera in real time and recover the 3D positions of thousands of points in a desktop-like environment. Due to the fact that a camera-pose update with a fixed map is much more efficient than a map update with known camera poses, Klein and Murray proposed to separate the tracking and mapping into two parallel threads. To achieve real-time speed, the tracking thread was given higher priority than the mapping thread (120). In the mapping thread, the time-consuming bundle adjustment optimization (121,122) was run to refine the stored 3D points and camera poses (120). The benefits of separating tracking and mapping include more robust camera tracking and more accurate 3D point positions.

Based on the results of camera tracking from PTAM, many research efforts (123–125) have been proposed to generate a consistent dense 3D model in real time. In (123), the 3D points from PTAM were triangulated to build a base mesh using Multi-Scale Compactly Supported Radial Basis Function (MSCSRBF) (126). This base mesh was then used to generate a synthesized image, which was compared with the real images captured by the camera at the same position to iteratively polish the dense model (123). In (123), the Total Variation regularization with L1 norm (TV-L1) optical flow (127) was applied to establish the correspondences between synthesized and real images. The dense model from (123) was later used to improve the camera tracking in (128). Instead of using variational optical flow, as in (123), Graber *et al*. (124) adopted the multiview plane-sweep to perform 3D reconstruction with high-quality depth map fusion (124). Based on PTAM and the work of Graber *et al*. (124), Wendel *et al*. (125) developed a live dense volumetric reconstruction for micro-aerial vehicles.

After the recovery of the 3D structure from monocular endoscopic videos, researchers have attempted to register the recovered 3D structures with the pre-operative data. Burschka *et al*. (105,106) proposed to register the recovered 3D points with a pre-operative CT model to achieve accurate navigation in sinus surgeries. To obtain accurate navigation for ESBS surgeries, Mirota *et al*. (129) introduced a new registration method, which was later applied (96,130) to register the 3D point cloud from (108) with the CT data.

Stereo cameras

Stereo cameras have gained popularity recently in robot-assisted surgery, such as using the da Vinci system (16,17). Compared with the bearing-only sensor of the monocular camera, stereo cameras can get the 3D locations of landmarks directly from a single measurement. In robotics, the early work of SLAM with stereo cameras on a mobile robot can be found in (131–133) and the more recent work is focused on how to apply stereo cameras in a large environment (134–138). Mountney *et al*. (139) extended the stereo SLAM technique in the MIS environment to track the stereoscope and reconstruct a sparse set of 3D points. A Shi–Tomasi feature point detector was used to find interest points, which were represented by 25 pixel x 25 pixel patches and tracked using ZSSD correlation (139). A 'constant velocity and constant angular velocity' model was adopted to describe the endoscope motion (139).

The stereo EKF–SLAM framework in (139) has been adopted in many different systems. Noonan *et al*. (140) applied the framework to track the newly-designed stereoscopic fibrescope imaging system. To get a larger field of view for surgeons, Mountney and Yang (139) integrated the output of the sparse 3D points and camera tracking results from (139) into the dynamic view expansion system (141) and textured the 3D mesh with past and current images (142). Warren *et al*. (143) pointed out that disorientation was a major challenge in natural orifice transluminal surgery (NOTES). They used an inertial measurement unit (IMU) attached at the tip of the endoscope to stabilize the image horizontally (143). The stabilized images were further integrated into the dynamic view expansion system (142) to provide more realistic navigation results (143). Totz *et al*. (72) reported that the sparse 3D mesh generated from the Mountney *et al*. stereo SLAM was not rich enough to represent the real 3D shape of the scene, which caused visual artifacts in the final textured-mapped 3D model. To overcome this problem, Totz *et al*. (72) used the sparse 3D points to register a couple of semi-dense 3D surfaces from stereo reconstruction (18) together to generate a larger and more accurate 3D model, which resulted in more consistent rendering results with dynamic view expansion.

Discussion

When enough texture is available on tissue surfaces, it has been shown that visual SLAM is able to estimate camera poses and recover a sparse set of 3D points with reasonable qualities (118,144). However, the results of visual SLAM depend greatly on the successful extraction of distinctive image features. Therefore, further studies are needed to extract distinctive image features for MIS images. On the other hand, a new visual SLAM framework was recently presented and no detection of image feature points was required (145,146). The system exploits and reconstructs each pixel with valid image gradients. This framework does not rely on image feature points and can be useful for the MIS environment.

Dynamic MIS–VSLAM

The assumption of rigid scenes in the previous section might not be valid in a general MIS environment. This section focuses on the general problem of MIS–VSLAM, which is termed 'dynamic MIS–VSLAM', as illustrated in Figure 7. In a typical MIS environment, the tissue surfaces might undergo non-rigid deformation caused by heartbeats, breathing and interaction with surgical instruments. Meanwhile, surgical instruments might move dynamically in the scene and cause occlusion problems. As a result, there are two fundamental tasks for dynamic MIS–VSLAM: the theoretical treatment of tissue deformation (16,17) and moving-instrument tracking.

The first task is similar to the recovery of surface deformation with a moving camera, which has been an active research topic in computer vision and belongs to the broader topic, non-rigid structure from motion (NRSFM) (147). NRSFM has been proposed to analyse non-rigid scenes, such as smooth surfaces, articulated bodies and piecewise rigid surfaces (148). The general problem of NRSFM is considered to be that ill-posed if arbitrary deformation is allowed (148). In the MIS environment, smooth tissue surfaces cast additional constraints on the general NRSFM and therefore the problem becomes less difficult. This section focuses on the introduction of the two essential tasks in dynamic MIS–VSLAM; NRSFM with deforming tissue surfaces, termed 'deforming surface SFM' (DSSFM), and dynamic surgical instrument tracking. Many approaches have been presented to tackle the problem of DSSFM and can be broadly classified into two categories, based on whether a monocular camera or stereo cameras are used. Those approaches are summarized in Table 2.

DSSFM with monocular cameras

Different from rigid SFM, each point in DSSFM can deform due to both global rigid motion and local deformation, which are difficult to differentiate. Therefore, different constraints of deformation from the inherent geometry of the shape have been introduced (147,149,150,154,158,159). It is generally considered that the work of Bregler *et al*. (149) was the first approach that successfully extended Tomasi and Kanades' factorization method (160) to non-rigid scenes. In (149), the idea was introduced of representing a 3D shape as a linear combination of a set of basis shapes, which greatly reduced the number of unknown parameters. This idea of a linear combination of basis shapes has been widely adopted since it was introduced. Most subsequent research has focused on convergence of the optimization by adding spatial and temporal smoothness constraints (154). One impractical assumption of the Bregler *et al*. method is the scaled orthographic camera model. This camera

DOI: 10.1002/rcs

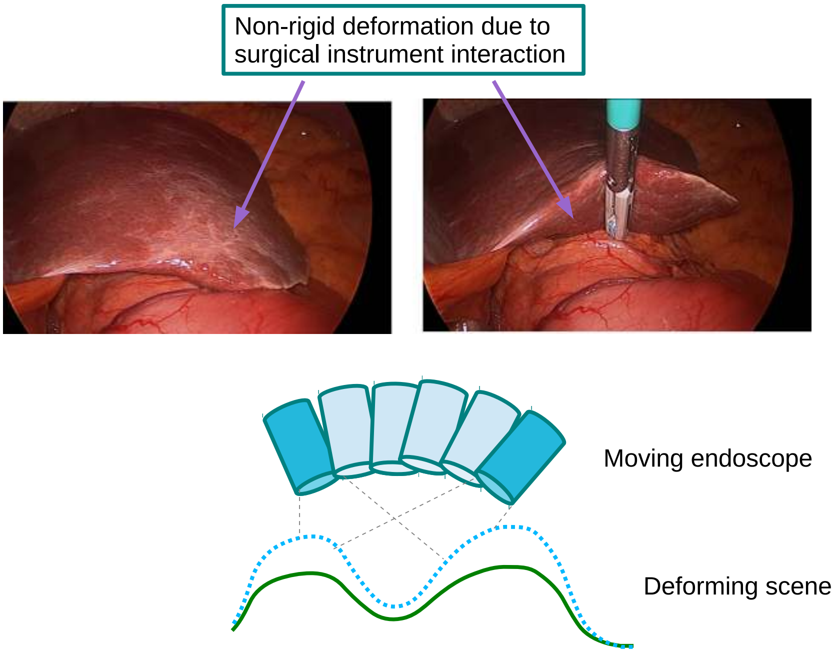
Non-rigid deformation due to surgical instrument interaction
Moving endoscope
Deforming scene

Figure 7. **MIS–VSLAM methods in dynamic or deforming scenes; solid green and dotted blue curves are used to illustrate non-rigid tissue deformation**

Table 2. **Summary of different approaches in DSSFM**

Methods	Sensor type	Batch/sequential	Rigid	Camera model
Bregler *et al*. (149)	Mono	Batch	No	Scaled orthographic
Xiao and Kanade (150)	Mono	Batch	No	Perspective
Del Bue *et al*. (151)	Mono	Batch	Yes	Perspective
Wang and Wu (152)	Mono	Batch	Yes	Affine
Hartley and Vidal (153)	Mono	Batch	No	Perspective
Paladini *et al*. (154)	Mono	Sequential	No	Orthographic
Del Bue and Agapito (155)	Stereo	Batch	No	Affine
Bartoli (156)	3D sensor	Batch	No	–
Llado (157)	Stereo	Batch	Yes	Perspective

Batch/sequential, optimization type; Rigid, assumption or not of rigid points.

model assumes that images are taken at a long distance from the objects. This restriction was later removed to allow the usage of more general-perspective cameras and obtain the closed-form solution for linear basis shape models (150,153).

Many DSSFM methods assume that all points are under non-rigid deformation, such as a piece of cloth under perturbation. In practice, a scene generally contains both rigid and non-rigid points – a common scenario in the MIS environment as well. In the publicly-available laparoscopic MIS image datasets (20), only those tissue organs that are interacting with surgical instruments display large deformation; other tissue surfaces mostly have small deformations that can sometimes be treated as rigid objects. Del Bue *et al*. (151) introduced the idea of assuming the existence of both rigid and non-rigid points. For a monocular camera, Del Bue *et al*. (151) used the RANSAC algorithm to segment rigid and non-rigid points, based on the criterion that only rigid points could satisfy epipolar geometry. The purpose of the Del Bue *et al*. method is to estimate the 3D shape of human faces, where there are many fewer rigid points than non-rigid ones. The small percentage of rigid points requires a large number of samplings in the RANSAC process, which greatly slows the segmentation. In (151), to speed up the RANSAC process, the degree of non-rigidity (DoN) was calculated for each point as in the prior information and was used to guide the sampling of RANSAC. DoN was defined based on the observation that 3D positions of non-rigid points change from time to time and, hence, have larger variances than the rigid ones (151).

One major challenge of many existing monocular DSSFM methods is the expensive time consumption of the final non-linear optimization. Motivated by the significant performance of PTAM (120), Paladini *et al.* (154) proposed the first work to separate model-based camera tracking and model updating. To enable sequential model updating, a sequential framework was presented that increased the degrees of freedom of basis shape whenever the current shape model was not able to represent a new shape (154). Based on dense 2D correspondences from (161), Garg *et al.* (148) formulated DSSFM as a variational energy minimization problem to estimate the 3D structure of deformable surface from a monocular video sequence.

DSSFM with stereo cameras

Since the relative pose between the stereo cameras was fixed, the factorization method (149) was extended to stereo cameras by stacking the constraints from each camera together (155,162). A novel method of decomposing the measurement matrix to get stereo camera pose and 3D shape was presented in (155,162). When corresponding points between stereo cameras are available, the 3D positions of those points for each frame can be obtained through triangulation. Therefore, the input to the DSSFM becomes 3D point tracks, rather than 2D point tracks as in the monocular case. With 3D point tracks as input, Llado *et al.* (157) extended the rigid and non-rigid point segmentations from a monocular camera to stereo cameras based on the fact that only rigid points satisfied a global Euclidean transformation. After RANSAC estimation, the classification of rigid points and non-rigid points was further refined, based on accumulated 3D registration errors, which were large for non-rigid points and small for rigid ones (157).

Table 3. Dynamic visual SLAM methods for MIS

Scene	Monocular camera	Stereo cameras
Rigid	(96,105–107,109,115,130)	(72,139,140,142,143,164)
Deforming	(116,165,166)	(144,167,168)

Another significant stereo DSSFM method was presented by Bartoli (156), who first learned the basis shapes by maximum likelihood, and then the learned basis shapes were used to estimate the stereo rig's poses as well as the configuration weights. There is a major difference between the Llado *et al.* method (157) and Bartoli methods (156). In (157), Llado *et al.* estimated the poses, basis shapes and configuration weights all together by non-linear optimization, which minimized the reprojection error. Bartoli (156) proposed to learn the basis shapes first from a sequence of 3D shapes, and then minimized the 3D registration error to estimate basis shapes and configuration weights. Besides stereo cameras, a multi-camera set-up (163) has also been considered to solve the DSSFM problem. Even though many methods have been presented, DSSFM is still considered a very difficult problem and remains an open challenge to researchers.

DSSFM in MIS environment

Currently, most DSSFM methods assume that all 3D points are correctly detected and tracked in each pair of stereo images. This assumption is generally not practical, because the feature matching might contain mismatches due to noise. In the MIS environment, low-contrast images, non-rigid deformation of organs and the dynamic moving of surgical instruments further complicate this issue. Despite these difficulties, different methods have been proposed to simplify the problem by adding practical constraints from the MIS environment, as summarized in Table 3. Some significant methods were chosen as representative; their properties are displayed in Table 4. In this section, we first introduce the methods proposed to overcome tissue deformation and then discuss methods designed to track moving objects.

Table 4. Summary of the state-of-the-art methods in MIS-VSLAM

Methods	Rigid/ deform	Batch/ sequ	Framework	Mono/stereo	Feature detection	Feature matching	Organs	Reg
Burschka *et al.* (106)	Rigid	Sequ	ASKC	Mono	Segmentation	SSD	Sinus	Yes
Wang *et al.* (107)	Rigid	Sequ	ASKC	Mono	SIFT + SVD	SIFT	Sinus	No
Mirota *et al.* (130)	Rigid	Sequ	ASKC	Mono	SIFT + SVD	SIFT	Sinus	Yes
Hu *et al.* (103)	Rigid	Batch	Factor	Mono/stereo	–	LK	Heart	Yes
Mountney *et al.* (139)	Rigid	Sequ	EKF–SLAM	Stereo	Shi–Tomasi	NSSD	Abdomen	No
Totz *et al.* (72)	Rigid	Sequ	EKF–SLAM	Stereo	Shi–Tomasi	NSSD	Abdomen	No
Hu *et al.* (166)	Deform	Batch	NRSFM	Mono	–	LK	Heart	No
Grasa *et al.* (116)	Deform	Sequ	EKF–SLAM	Mono	FAST	NSSD	Abdomen	No
Collins *et al.* (165)	Deform	Batch	–	Mono	–	Optical flow (169)	Liver	No
Mountney and Yang (167)	Deform	Sequ	EKF–SLAM	Stereo	Shi–Tomasi	(49)	Abdomen	No
Lin *et al.* (144)	Deform	Sequ	PTAM (120)	Stereo	FAST	ZSSD	Abdomen	No

Rigid/deform, rigid or deforming scene; Batch/sequ, batch or sequential optimization; Framework, type of optimization framework; Mono/stereo, monocular camera or stereo cameras; Feature detection, types of feature detection; Feature matching, types of feature matching; Reg, registration; Sequ., sequential; ASKC, optimization method proposed in (107); Factor, matrix factorization method of rigid SFM; NSSD, normalized SSD.

DOI: 10.1002/rcs

Many researchers have attempted to reduce tissue deformation by rearranging or segmenting the videos. Hu *et al*. (166) applied the probabilistic principal component analysis (PPCA)-based NRSFM (147) to reconstruct a beating heart surface and estimate the camera poses. To reduce complexity from deformation, the video sequence was rearranged, and the images of the same heart cycles were chosen to reduce tissue deformation (166); in this method some feature points may be lost during the tracking, and it is unclear how this problem is compensated for. Collins *et al*. (165) argued that tissue motion was small within a couple of frames and, hence, could be treated as rigid. With this assumption, Collins *et al*. (165) presented a method to divide the video sequence into small segments, and the motion within each segment was approximated as rigid.

Researchers also observed that the deformation of particular organs, such as the liver, might follow certain periodic patterns, such as respiration and heartbeats. These periodic patterns can be learned and used as constraints to overcome the challenges from tissue deformation. Mountney *et al*. (167) presented a SLAM framework for the MIS environment with periodic tissue deformation. In (167), liver motion was described by a periodic respiration model and learned by temporally tracking the 3D points on the liver surface, using stereo cameras. The learned respiration model was later integrated into the EKF framework for more accurate prediction of camera poses. However, the assumption of periodic motion is not valid for all tissues; for example, the tissue motion caused by the interaction of surgical instruments is mostly not periodic.

PTAM has been shown to be robust in a desktop environment, and its application in the MIS environment is not as stable, because of difficulties from the less distinctive features in MIS images and non-rigid tissue deformation. Lin *et al*. (144) extended monocular PTAM (120) to a stereoscope and proposed to use RANSAC to detect the deforming points, based on the fact that only rigid points satisfy a global Euclidean transformation. The removal of the deforming points resulted in more accurate and stable camera pose estimation results. When the stereoscope is available, there is no need for manual initialization, as in (120), and the scale of the recovered structure can be determined. The stereoscope PTAM was applied in a laparoscopic video, and a bladder model was highlighted in the video to remind the surgeons (144). Figure 8 shows typical frames extracted from a MIS video with an overlaid bladder model.

With the development and availability of miniaturized microelectromechanical systems, researchers have been trying to use inertial sensors to further improve visual SLAM performance. Giannarou *et al*. (170) presented a novel method, adaptive unscented Kalman filter (UKF), to exploit the data from an IMU (170). The IMU data were combined with visual information to achieve better camera pose estimation for deformable scenes in MIS (170).

Moving instrument tracking

In visual SLAM, dynamic moving objects usually result in inaccurate camera localization results. Therefore, it is necessary to track these moving objects. In robotics, this problem is generally referred to as 'SLAM and moving object tracking' (SLAMMOT), which deals with dynamic environments containing moving objects, such as humans and cars. Wang and Thorpe (171) reported the first work that successfully detected and tracked moving objects within a visual SLAM system. A mathematical framework was introduced (172) and a general solution was provided to the problem of SLAMMOT. Recently, Lin and Wang (173) presented a stereo camera-based approach for SLAMMOT, which overcame the observability issue that was common in monocular approaches. Zou *et al*. (174) presented the first work that applied visual SLAM using multiple independent cameras in a dynamic environment, in which it was shown that with multiple cameras, the rigid and moving points could be distinguished based on the reprojection distance. Also, each camera's pose and the 3D locations of moving points could be successfully recovered by considering nearby cameras' observations of the landmarks (174).

In MIS, different techniques have been introduced to track the dynamically moving surgical instruments. In a typical MIS set-up, the instruments are inserted through

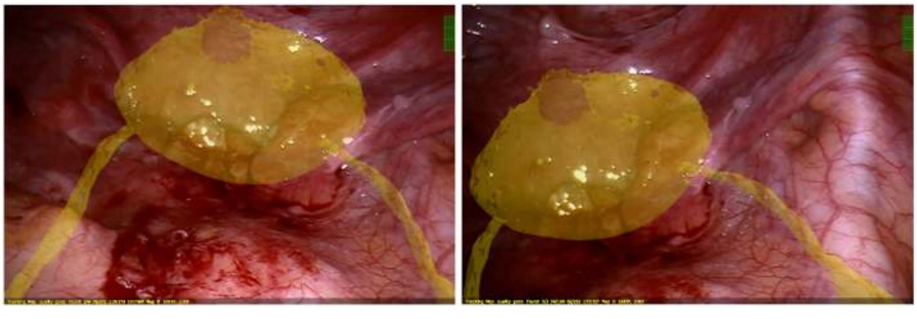

Figure 8. Example of a bladder model (yellow) overlaid on a MIS video (144)

DOI: 10.1002/rcs

small incisions and their motions are, therefore, greatly restricted. Voros *et al.* measured the 3D position of the insertion point of an instrument and exploited the 3D instrument model to constrain the search space and achieve accurate instrument detection (175,176). Allan *et al.* (24) argued that the estimated trocar positions might be inaccurate, due to trocar and patient movement. They proposed a probabilistic supervised classification method, which did not require the estimation of the trocar positions (24). Instead, they first detected surgical-instrument pixels and then estimated the poses of those instruments (24).

Endoscope video-based object tracking has many applications. In (177) a suturing needle was tracked, and 3D cue information was augmented in the video to help surgeons better understand the poses of the needle. Jayarathne *et al.* (178) introduced a method to track the ultrasound probe using the standard monocular endoscopic camera, so that magnetic tracking could be obviated. They presented an EKF framework to establish the correspondences and estimated the pose of the ultrasound probe (178).

Discussion

To overcome the difficulties in dynamic MIS–VSLAM, it is essential to exploit the prior information of surgical scenes and use them as constraints. Since tissue surfaces are smooth and have special deforming properties, one important research topic is to learn biomechanical models of tissue deformation. Organs and tissues have specific shapes and biological properties, which greatly restrict how they would deform. Those biomechanical models are usually similar among different people and can be learned before the surgery.

In the abdominal environment, large areas of surgical scenes, such as abdominal walls, typically remain relatively still during the whole surgical procedure. These rigid areas can be pre-identified and used to separate camera-pose estimation and deformation recovery. 3D models of surgical instruments can be exploited as prior information to assist instrument tracking.

Video-based camera localization and 3D reconstruction rely on robust image feature detection and matching results. However, some tissue surfaces do not have distinctive textures, e.g. the texture of the liver surface is repetitive and indistinctive. In those scenarios, extra information from tissue organs is necessary. For instance, the contours of a liver can be accurately detected and matched to its 3D model from preoperative data to estimate its pose and deformation. Another option is to actively project patterns on tissue surfaces to build corresponding points for 3D reconstruction. Therefore, structured lighting-based methods, such as depth sensors, are important to solve the low-texture problem.

Results

Videos captured *in situ* during MIS have enabled the use of vision-based techniques to assist surgeons to better visualize a surgical site and navigate inside a body. Video-based surgical-scene 3D reconstruction, laparoscope localization and deformation recovery are fundamental problems for dynamic registration and surgical navigation in abdominal MIS. This paper has reviewed the methods of feature detection and tracking for MIS images. Additionally, this paper has summarized 3D reconstruction and visual SLAM methods for rigid surgical scenes in MIS. Moreover, this paper has introduced methods for deformation recovery and summarized 3D reconstruction and visual SLAM for deforming tissue organs.

Multiple results have been obtained. The publicly available datasets have been collected in Table 1. The state-of-the-art DSSFM methods have been provided in Table 2. The 3D reconstruction, laparoscope localization and deformation recovery techniques for dynamic surgical scenes with general tissue deformation and instrument occlusions have been summarized in Tables 3 and 4.

Even though much research work has been presented in the field, simultaneous 3D reconstruction, laparoscope localization and deformation recovery in real time for a dynamic MIS environment is still difficult and remains an open challenge. To achieve this goal, multiple important research directions need further exploration. First, many well-studied computer vision techniques, such as 3D reconstruction and camera localization, are based on the successful detection of distinctive image features. It has been known that the MIS environment is quite different from the man-made environment, and it is desirable to analyse and exploit the special characteristics of MIS images. Some recent research (12,22,38,59) along this direction has been made available; however, more research in this area is needed to fundamentally solve the problem.

Multiple visual SLAM frameworks have been presented in the literature, such as PTAM (120,144) and EKF–SLAM (15,139). These frameworks have been validated to work well when surgical scenes are mostly rigid and camera motions are not too fast. To make endoscope localization more robust, it is necessary to extend the SLAM system to work in a deforming environment. However, if any arbitrary deformation is allowed, the problem becomes infeasible because of the ambiguity of distinguishing between camera movement and scene deformation. Therefore, the key to solving this problem is to carefully discover and learn the deformation models. One research direction is to assume that only certain parts of a scene are deforming and the others are rigid (144). Since organ deformation follows biomechanical properties, another promising research direction is to learn those biomechanical models.

Medical data are crucial for the research community. MIS data acquisition is known to be difficult (86) and has become one of the main challenges in this area. As listed in Table 1, multiple MIS datasets have recently become available to the public. However, the abdominal environment is complex and more MIS datasets are still needed. Meanwhile, standardized *in vivo* evaluation methods/procedures are also necessary in order to compare different research studies or reproduce other research work.

Conflict of interest

The authors have stated explicitly that there are no conflict of interest in connection with this article.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. 1035594.

References

1. Sun Y, Anderson A, Castro C, *et al*. Virtually transparent epidermal imagery for laparo-endoscopic single-site surgery. In International Conference of the IEEE Engineering in Medicine and Biology Society, 2011; 2107–2110.
2. Anderson A, Lin B, Sun Y. Virtually transparent epidermal imagery (VTEI): on new approaches to *in vivo* wireless high-definition video and image processing. *IEEE Trans Biomed Circuits Syst* 2013; **99**: 1–1.
3. Maintz J, Viergever MA. A survey of medical image registration. *Med Image Anal* 1998; **2**: 1–36.
4. Rueckert D, Schnabel J. Medical image registration. In *Biomedical Image Processing*. Series: Biological and Medical Physics; Biomedical Engineering. Springer: Berlin, Heidelberg, 2011; 131–154.
5. Markelj P, Tomazevic D, Likar B, *et al.* A review of 3D/2D registration methods for image-guided interventions. *Med Image Anal* 2012; **16**: 642–661.
6. Zitova B, Flusser J. Image registration methods: a survey. *Image Vis Comput* 2003; **21**: 977–1000.
7. Sotiras A, Davatzikos C, Paragios N. Deformable medical image registration: a survey. *IEEE Trans Med Imag* 2013; **32**: 1153–1190.
8. Glocker B, Sotiras A, Komodaki N, *et al.* Deformable medical image registration: setting the state of the art with discrete methods. *Annu Rev Biomed Eng* 2011; **13**: 219–244.
9. Grimson E, Leventon M, Ettinger G, *et al.* Clinical experience with a high precision image-guided neurosurgery system. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*. Series: Lecture Notes in Computer Science, Vol. **1496**. Springer: Berlin, Heidelberg, 1998; 63–73.
10. Shamir R, Joskowicz L, Spektor S, *et al.* Localization and registration accuracy in image guided neurosurgery: a clinical study. *Int J Comput Assist Radiol Surg* 2009; **4**: 45–52.
11. Lin B, Sun Y, Qian X. Dense surface reconstruction with shadows in MIS. *IEEE Trans Biomed Eng* 2013; **60**: 2411–2420.
12. Lin B, Sun Y, Sanchez J, *et al*. Vesselness based feature extraction for endoscopic image analysis. In Proceedings of the International Symposium on Biomedical Imaging, 2014; 1295–1298.
13. Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: Part 1. *IEEE Robot Autom Mag* 2006; **13**: 99–110.
14. Bailey T, Durrant-Whyte H. Simultaneous localisation and mapping (SLAM): Part II. *State of the art. IEEE Robotics Autom Mag* 2006; **13**: 108–117.
15. Davison AJ. Real-time simultaneous localisation and mapping with a single camera. In Proceeding of the International Conference on Computer Vision, 2003; 1403–1410.
16. Mountney P, Stoyanov D, Yang GZ. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Process Mag* 2010; **27**: 14–24.
17. Stoyanov D. Surgical vision. *Ann Biomed Eng* 2012; **40**: 332–345.
18. Stoyanov D, Scarzanella MV, Pratt P, *et al*. Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2010; 275–282.
19. Bouguet JY. Camera calibration toolbox: http: //www.vision.caltech.edu/bouguetj/calib doc/index.html#ref
20. Giannarou S, Stoyanov D, Noonan D, *et al*. 2014; Hamlyn centre laparoscopic/endoscopic video datasets: http://hamlyn.doc.ic.ac.uk/vision/
21. Bartoli A. 2014. Monocular laparoscopic video dataset of uterus: http://isit.u-clermont1.fr/∼ab/Research/Datasets/Uterus 01.rar
22. Lin B, Sun Y, Sanchez J, *et al*. Efficient vessel feature detection for endoscopic image analysis. *IEEE Trans Biomed. Imag* 2014: http://rpal.cse.usf.edu/project1/index.html
23. Puerto-Souza G, Mariottini GL. A fast and accurate feature-matching algorithm for minimally-invasive endoscopic images. *IEEE Trans Med Imag* 2013; **32**: 1201–1214: http://ranger.uta.edu/∼gianluca/feature matching/
24. Allan M, Ourselin S, Thompson S, *et al*. Toward detection and localization of instruments in minimally invasive surgery. *IEEE Trans Biomed Eng* 2013; **60**: 1050–1058: http://www.surgicalvision.cs.ucl.ac.uk/benchmarking/#home
25. Gao Y, Vedula SS, Reiley CE, *et al*. The jhu-isi gesture and skill assessment dataset (jigsaws): a surgical activity working set for human motion modeling. In Medical Image Computing and Computer-Assisted Intervention (MICCAI) M2CAI Workshop, 2014: http://cirl.lcsr.jhu.edu/research/hmm/datasets/jigsaws release/
26. Speidel S, Kenngott H, Maier-Hein L. 2014; Open-CAS: validating and benchmarking computer assisted surgery: http://opencas.webarchiv.kit.edu/
27. Maier-Hein L, Groch A, Bartoli A, *et al.* Comparative validation of single-shot optical techniques for laparoscopic 3D surface reconstruction. *IEEE Trans Med Imaging* 2014; **33**(10): 1913–1930.
28. Allan M, Thompson SS, Clarkson M, *et al*. 2D–3D pose tracking of rigid instruments in minimally invasive surgery. In International Conference on Information Processing in Computer-assisted Interventions, 2014; **33**(10): 1913–1930.
29. Hoover A, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans Med Imag* 2000; **19**: 203–210.
30. Staal J, Abrmoff MD, Niemeijer M, *et al.* Ridge-based vessel segmentation in color images of the retina. *IEEE Trans Med Imag* 2004; **23**: 501–509.
31. Al-Diri B, Hunter A, Steel D, *et al*. Review – a reference data set for retinal vessel profiles. In International Conference of the IEEE Engineering in Medicine and Biology Society, 2008; 2262–2265.
32. Rosten E, Drummond T. Machine learning for high-speed corner detection. In Proceedings of the European Conference on Computer Vision 1, May 2006; 430–443.
33. Mair E, Hager G, Burschka D, *et al*. Adaptive and generic corner detection based on the accelerated segment test. In

Proceedings of the European Conference on Computer Vision, 2010: **6312**; 183–196.

34. Harris C, Stephens M. A combined corner and edge detector. In Proceedings of the Alvey Vision Conference, 1988; 147–151.
35. Shi J, Tomasi C. Good features to track. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 1994; 593–600.
36. Mikolajczyk K, Schmid C. Scale and affine invariant interest point detectors. *Int J Comput Vis* 2004; **60**: 63–86.
37. Giannarou S, Visentini-Scarzanella M, Yang GZ. Affine-invariant anisotropic detector for soft tissue tracking in minimally invasive surgery. In Proceedings of the International Symposium on Biomedical Imaging, 2009; 1059–1062.
38. Giannarou S, Visentini Scarzanella M, Yang GZ. Probabilistic tracking of affine-invariant anisotropic regions. *IEEE Trans Pattern Anal Mach Intell* 2013; **35**: 130–143.
39. Mikolajczyk K, Tuytelaars T, Schmid C, *et al.* A comparison of affine region detectors. *Int J Comput Vis* 2005; **65**: 43–72.
40. Lowe DG. Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 2004; **60**: 91–110.
41. Bay H, Tuytelaars T. Gool LV. Surf: Speeded up robust features. In Proceedings of the European Conference on Computer Vision, 2006; 404–417.
42. Stavens D, Thrun S. Unsupervised learning of invariant features using video. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2010; 1649–1656.
43. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. In Proceedings of the International Joint Conference on Artificial Intelligence, 1981; 674–679.
44. Lepetit V, Pilet J, Fua P. Point matching as a classification problem for fast and robust object pose estimation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2004; 244–250.
45. Lepetit V, Lagger P, Fua P. Randomized trees for real-time keypoint recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2005; 775–781.
46. Lepetit V, Fua P. Keypoint recognition using randomized trees. *IEEE Trans Pattern Anal Mach Intell* 2006; **28**: 1465–1479.
47. Ozuysal M, Fua P, Lepetit V. Fast keypoint recognition in ten lines of code. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2007.
48. Ozuysal M, Calonder M, Lepetit V, *et al.* Fast keypoint recognition using random ferns. *IEEE Trans Pattern Anal Mach Intell* 2010; **32**: 448–461.
49. Mountney P, Yang GZ. Soft tissue tracking for minimally invasive surgery: learning local deformation online. Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2008; **11**: 364–372.
50. Lim J, Yang MH. A direct method for modeling non-rigid motion with thin plate spline. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2005; 1196–1202.
51. Richa R, Poignet P, Liu C. Efficient 3D tracking for motion compensation in beating heart surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2008; 684–691.
52. Stoyanov D, Mylonas G, Deligianni F, *et al*. Soft tissue motion tracking and structure estimation for robotic assisted mis procedures. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, Duncan J, Gerig G (eds). Series: Lecture Notes in Computer Science. Springer: Berlin, Heidelberg, 2005; **3750**: 139–146.
53. Matas J, Chum O, Urban M, *et al. Robust wide baseline stereo from maximally stable extremal regions*. : In Proceedings of the British Machine Vision Conference, 2002.
54. Hu M, Penney GP, Edwards PJ, *et al*. 3D reconstruction of internal organ surfaces for minimal invasive surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2007; 68–77.
55. Richa R, Bo APL, Poignet P. Robust 3D visual tracking for robotic-assisted cardiac interventions. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2010; 267–274.
56. Calonder M, Lepetit V, Strecha C, *et al*. Binary robust independent elementary features. In Proceedings of the European Conference on Computer Vision: Brief, 2010; 778–792.
57. Puerto Souza GA, Adibi M, Cadeddu JA, *et al*. Adaptive multiaffine (AMA) feature-matching algorithm and its application to minimally-invasive surgery images. In Intelligent Robots and Systems, 2011; 2371–2376.
58. Puerto-Souza G, Mariottini G. Hierarchical multi-affine (HMA) algorithm for fast and accurate feature matching in minimally-invasive surgical images. In Proceedings of the IEEE/RSJ International Intelligent Robots and Systems Conference, 2012; 2007–2012.
59. Lin B, Sun Y, Qian X. Thin plate spline feature point matching for organ surfaces in minimally invasive surgery imaging. In SPIE Medical imaging, 2013.
60. Mountney P, Lo BPL, Thiemjarus S, *et al*. A probabilistic framework for tracking deformable soft tissue in minimally invasive surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2007; 34–41.
61. Yip MC, Lowe DG, Salcudean SE, *et al.* Real-time methods for long-term tissue feature tracking in endoscopic scenes. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*. Series: Lecture Notes in Computer Science, Vol. **7330**. Springer: Berlin, Heidelberg, 2012; 33–43.
62. Yip MC, Lowe DG, Salcudean SE, *et al.* Tissue tracking and registration for image-guided surgery. *IEEE Trans Med Imag* 2012; **31**: 2169–2182.
63. Garage W. Star detector: http://pr.willowgarage.com/wiki/ Star Detector
64. Agrawal M, Konolige K. Blas MR. Censure: Center surround extremas for realtime feature detection and matching. In Proceedings of the European Conference on Computer Vision, 2008; 102–115.
65. Maier-Hein L, Mersmann S, Kondermann D, *et al*. Can masses of non-experts train highly accurate image classifiers? In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*. Series: Lecture Notes in Computer Science. Springer: Berlin, Heidelberg, 2014; **8674**: 438–445.
66. Maier-Hein L, Mersmann S, Kondermann D, *et al.* Crowd sourcing for reference correspondence generation in endoscopic images. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*. Series: Lecture Notes in Computer Science, Vol. **8674**. Springer: Berlin, Heidelberg, 2014; 349–356.
67. Klippenstein J, Zhang H. Quantitative evaluation of feature extractors for visual slam. In Computer and Robot Vision, 2007; 157–164.
68. Selka F, Nicolau S, Agnus V, *et al.*Evaluation of endoscopic image enhancement for feature tracking: a new validation framework. In *Medical Imaging and Augmented Reality*. Springer: Berlin, Heidelberg, 2013; **8090**; 75–85.
69. Lau WW, Ramey NA, Corso JJ, *et al.* Stereo-based endoscopic tracking of cardiac surface deformation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*. Springer: Berlin, Heidelberg, 2004; 494–501.
70. Kowalczuk J, Meyer A, Carlson J, *et al.* Real-time three-dimensional soft tissue reconstruction for laparoscopic surgery. *Surg Endosc* 2012; **26**: 3413–3417.
71. Totz J, Thompson S, Stoyanov D, *et al*. Fast semi-dense surface reconstruction from stereoscopic video in laparoscopic surgery. In *Information Processing in Computer-Assisted Interventions*. Series: Lecture Notes in Computer Science. Springer: Berlin, Heidelburg, 2014; **8498**; 206–215.

72. Totz J, Mountney P, Stoyanov D, *et al*. Dense surface reconstruction for enhanced navigation in MIS. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2011; 89–96.
73. Scarzanella MV, Mylonas GP, Stoyanov D, *et al*. i-brush: a gaze-contingent virtual paintbrush for dense 3D reconstruction in robotic assisted surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2009; 353–360.
74. Bernhardt S, Abi-Nahed J, Abugharbieh R. Robust dense endoscopic stereo reconstruction for minimally invasive surgery. In *Medical Computer Vision Recognition Techniques and Applications in Medical Imaging*, Vol. **7766**. Springer: Berlin, Heidelberg, 2012; 254–262.
75. Chang PL, Stoyanov D, Davison A, *et al.* Real-time dense stereo reconstruction using convex optimisation with a cost-volume for image-guided robotic surgery. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, Vol. **8149**. Springer: Berlin, Heidelberg, 2013; 42–49.
76. McKinlay R, Shaw M, Park A. A technique for real-time digital measurements in laparoscopic surgery. *Surg Endosc* 2004; **18**: 709–712.
77. Hayashibe M, Suzuki N, Nakamura Y. Laser-scan endoscope system for intraoperative geometry acquisition and surgical robot safety management. *Med Image Anal* 2006; **10**: 509–519.
78. Koninckx TP, Van Gool L. Real-time range acquisition by adaptive structured light. *IEEE Trans Pattern Anal Mach Intell* 2006; **28**: 432–445.
79. Salvi J, Fernandez S, Pribanic T, *et al.* A state of the art in structured light patterns for surface profilometry. *Pattern Recogn* 2010; **43**: 2666–2680.
80. Albitar C, Graebling P, Doignon C. Robust structured light coding for 3D reconstruction. In Proceedings of the Conference on Computer Vision, 2007; 1–6.
81. Maurice X, Graebling P, Doignon C. A pattern framework driven by the hamming distance for structured light-based reconstruction with a single image. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2011; 2497–2504.
82. Schmalz C, Forster F, Schick A, *et al.* An endoscopic 3D scanner based on structured light. *Med Image Anal* 2012; **16**: 1063–1072.
83. Clancy NT, Stoyanov D, Maier-Hein L, *et al.* Spectrally encoded fiber-based structured lighting probe for intraoperative 3D imaging. *Biomed Opt Express* 2011; **2**: 3119–3128.
84. Penne J, Höller KS, *et al*. Time-of-flight 3D endoscopy. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2009; 467–474.
85. Haase S, Bauer S, Wasza J, *et al.* 3D operation situs reconstruction with time-of-flight satellite cameras using photogeometric data fusion. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, Vol. **8149**. Springer: Berlin, Heidelberg, 2013; 356–363.
86. Maier-Hein L, Mountney P, Bartoli A, *et al.* Optical techniques for 3D surface reconstruction in computer-assisted laparoscopic surgery. *Med Image Anal* 2013; **17**: 974–996.
87. Ciuti G, Visentini-Scarzanella M, Dore A, *et al.* Intra-operative monocular 3D reconstruction for image guided navigation in active locomotion capsule endoscopy. In Proceedings of the 4th IEEE RAS and EMBS International Biomedical Robotics and Biomechatronics (BioRob) Conference, 2012; 768–774.
88. Wu C, Narasimhan S, Jaramaz B. A multi-image shape from shading framework for near-lighting perspective endoscopes. *Int J Comput Vis* 2010; **86**: 211–228.
89. Woodham RJ. Photometric method for determining surface orientation from multiple images. *Opt Eng* 1980; **19**: 139–144.
90. Zhang R, Tsai PS, Cryer JE, *et al.* Shape from shading: a survey. *IEEE Trans Pattern Anal Mach Intell* 1999; **21**: 690–706.
91. Mishra RK, Hanna GB, Brown SI, *et al.* Optimum shadow-casting illumination for endoscopic task performance. *Arch Surg* 2004; **139**: 889–892.
92. Nicolaou M, James A, Lo BPL, *et al*. Invisible shadow for navigation and planning in minimal invasive surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2005; 25–32.
93. Bouguet JY, Perona P. 3D photography using shadows in dual-space geometry. *Int J Comput Vis* 1999; **35**: 129–149.
94. Stoyanov D. Stereoscopic scene flow for robotic assisted minimally invasive surgery. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*. Series: Lecture Notes in Computer Science, Vol. **7510**. Springer: Berlin, Heidelberg, 2012; 479–486.
95. Chadebecq F, Tilmant C, Bartoli A. How big is this neoplasia? Live colonoscopic size measurement using the in-focus breakpoint. *Med Image Anal* 2015; **19**: 58–74.
96. Mirota D, Wang H, Taylor RH, *et al.* Toward video-based navigation for endoscopic endonasal skull base surgery. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, Vol. **12**. Springer: Berlin, Heidelberg, 2009; 91–99.
97. Shahidi R, Bax M, Maurer J, *et al.* Implementation, calibration and accuracy testing of an image-enhanced endoscopy system. *IEEE Trans Med Imaging* 2002; **21**: 1524–1535.
98. Lapeer R, Chen MS, Gonzalez G, *et al.* Image enhanced surgical navigation for endoscopic sinus surgery: evaluating calibration, registration and tracking. *Int J Med Robotics Comput Assist Surg* 2008; **4**: 32–45.
99. Furukawa Y, Ponce J. Accurate, dense, and robust multi-view stereopsis. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2007; 1–8.
100. Furukawa Y, Curless B, Seitz SM, *et al*. Towards internetscale multi-view stereo. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2010; 1434–1441.
101. Wu CH, Sun YN, Chang CC. Three-dimensional modeling from endoscopic video using geometric constraints via feature positioning. *IEEE Trans Biomed Eng* 2007; **54**: 1199–1211.
102. Atasoy S, Noonan DP, Benhimane S, *et al*. A global approach for automatic fibroscopic video mosaicing in minimally invasive diagnosis. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2008; 850–857.
103. Hu M, Penney GP, Figl M, *et al.* Reconstruction of a 3D surface from video that is robust to missing data and outliers: application to minimally invasive surgery using stereo and mono endoscopes. *Med Image Anal* 2012; **16**: 597–611.
104. Sun D, Liu J, Linte C, *et al.* Surface reconstruction from tracked endoscopic video using the structure from motion approach. In *Medical Imaging and Augmented Reality*. Springer: Berlin, Heidelberg, 2013; 127–135.
105. Burschka D, Li M, Taylor RH, *et al*. Scale-invariant registration of monocular endoscopic images to CT-scans for sinus surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2004; 413–421.
106. Burschka D, Li M, Ishii M, *et al.* Scale invariant registration of monocular endoscopic images to CT-scans for sinus surgery. *Med Image Anal* 2005; **9**: 413–426.
107. Wang H, Mirota D, Ishii M, *et al*. Robust motion estimation and structure recovery from endoscopic image sequences with an adaptive scale kernel consensus estimator. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2008.
108. Delponte E, Isgrò F, Odone F, *et al.* SVD-matching using SIFT features. *Graph Models* 2006; **68**: 415–431.
109. Mori K, Deguchi D, Sugiyama J, *et al.* Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images. *Med Image Anal* 2002; **6**: 321–336.

DOI: 10.1002/rcs

110. Davison AJ, Reid ID, Molton ND, *et al.* Monoslam: real-time single camera slam. *IEEE Trans Pattern Anal Mach Intell* 2007; **29**: 1052–1067.
111. Montiel JMM, Civera J, Davison AJ, Unified inverse depth parametrization for monocular slam. In Robotics: Science and Systems, 2006.
112. Civera J, Davison A, Montiel JMM. Inverse depth parametrization for monocular slam. *IEEE Trans Robotics* 2008; **24**: 932–945.
113. Civera J, Grasa OG, Davison A, *et al*. One-point RANSAC for EKF-based structure from motion. In Proceedings of the International Conference on Intelligent Robots and Systems, 2009; 3498–3504.
114. Civera J, Grasa O, Davison A, *et al.* One-point RANSAC for extended Kalman filtering: application to real-time structure from motion and visual odometry. *J Field Robot* 2010; **27**: 609–631.
115. Grasa OG, Civera J, Guemes A, *et al*. EKF monocular slam 3D modeling, measuring and augmented reality from endoscope image sequences. In Workshop on Augmented Environments for Medical Imaging including Augmented Reality in Computer-Aided Surgery, 2009.
116. Grasa OG, Civera J, Montiel JMM. EKF monocular slam with relocalization for laparoscopic sequences. In Proceedings of the International Conference on Robotics and Automation, 2011; 4816–4821.
117. Williams BP, Klein G, Reid ID. Real-time slam relocalization. In Proceedings of the International Conference on Computer Vision, 2007; 1–8.
118. Grasa GO, Bernal E, Casado S, *et al.* Visual slam for hand-held monocular endoscope. *IEEE Trans Med Imag* 2013; **99**: 1–1.
119. Mouragnon E, Lhuillier M, Dhome M, *et al*. Real-time localization and 3D reconstruction. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2006; 363–370.
120. Klein G, Murray DW. Parallel tracking and mapping for small AR workspaces. In International Symposium on Mixed Augmented Reality, 2007; 225–234.
121. Hartley RI, Zisserman A. *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press: Cambridge, UK, 2004; ISBN: 0521540518.
122. Szeliski R. *Computer Vision: Algorithms and Applications*. Series: Texts in Computer Science. Springer: New York, London, 2010: http://opac.inria.fr/record=b1130924
123. Newcombe RA, Davison AJ. Live dense reconstruction with a single moving camera. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2010; 1498–1505.
124. Graber G, Pock T, Bischof H. Online 3D reconstruction using convex optimization. In IEEE International Conference on Computer Vision Workshops, 2011; 708–711.
125. Wendel A, Maurer M, Graber G, *et al*. Dense reconstruction on-the-fly. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2012; 1450–1457.
126. Ohtake Y, Belyaev AG, Seidel HP. A multi-scale approach to 3D scattered data interpolation with compactly supported basis function. *In Shape Modeling International* 2003; **153–164**: 292.
127. Zach C, Pock T, Bischof H. A duality based approach for realtime TV-L1 optical flow. In *Annual Symposium of the German Association for Pattern Recognition*. Springer: Berlin, Heidelberg, 2007; 214–223.
128. Newcombe R, Lovegrove S, Davison A. DTAM: Dense tracking and mapping in real time. In Proceedings of the International Conference on Computer Vision, 2011; 2320–2327.
129. Mirota D, Taylor RH, Ishii M, *et al*. Direct endoscopic video registration for sinus surgery. In Proceedings of SPIE Medical Imaging: Vision, Image-guided Procedures and Modeling, 2009; **7261**.
130. Mirota D, Wang H, Taylor RH, *et al.* A system for video-based navigation for endoscopic endonasal skull base surgery. *IEEE Trans Med Imag* 2012; **31**: 963–976.
131. Se S, Lowe DG, Little JJ. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. *Int J Robotic Res* 2002; **21**: 735–760.
132. Davison AJ. Mobile robot navigation using active vision PhD Dissertation, University of Oxford, 1998.
133. Davison AJ, Murray DW. Mobile robot localisation using active vision. *In Proceedings of the European Conference on Computer Vision*, 1998; 809–825.
134. Lemaire T, Berger C, Jung IK, *et al.* Vision-based slam: Stereo and monocular approaches. *Int J Comput Vis* 2007; **74**: 343–364.
135. Konolige K, Agrawal M. Frameslam: from bundle adjustment to real-time visual mapping. *IEEE Trans Robotics* 2008; **24**: 1066–1077.
136. Mei C, Sibley G, Cummins M, *et al.* RSLAM: a system for large-scale mapping in constant time using stereo. *Int J Comput Vis* 2001; **94**: 198–214.
137. Lim J, Frahm JM, Pollefeys M. Online environment mapping, in Proceedings of the Conference on Computer Vision and Pattern Recognition, 2011; 3489–3496.
138. Strasdat H, Davison AJ, Montiel JMM, *et al*. Double window optimisation for constant time visual slam. In Proceedings of the International Conference on Computer Vision, 2011; 2352–2359.
139. Mountney P, Stoyanov D, Davison AJ, *et al*. Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2006; 347–354.
140. Noonan DP, Mountney P, Elson DS, *et al*. A stereoscopic fibrescope for camera motion and 3D depth recovery during minimally invasive surgery. In Proceedings of the International Conference on Robotics and Automation, 2009; 4463–4468.
141. Lerotic M, Chung AJ, Clark J, *et al*. Dynamic view expansion for enhanced navigation in natural orifice transluminal endoscopic surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2008; 467–475.
142. Mountney P, Yang GZ. Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009; 1184–1187.
143. Warren A, Mountney P, Noonan D, *et al.* Horizon stabilized dynamic view expansion for robotic assisted surgery (HS-DVE). *Int J Computer Assist Radiol Surg* 2012; **7**: 281–288.
144. Lin B, Johnson A, Qian X, *et al.* Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery. *In Medical Imaging and Augmented Reality* 2013; **8090**: 35–44.
145. Engel J, Sturm J, Cremers D. Semi-dense visual odometry for a monocular camera. In International Conference on Computer Vision, 2013.
146. Engel J, Schöps T, *Cremers D*. LSD-SLAM: Large-scale direct monocular SLAM. In European Conference on Computer Vision, 2014.
147. Torresani L, Hertzmann A, Bregler C. Non-rigid structure from motion: estimating shape and motion with hierarchical priors. *IEEE Trans Pattern Anal Mach Intell* 2008; **30**: 878–892.
148. Garg R, Roussos A, Agapito L. Dense variational reconstruction of non-rigid surfaces from monocular video. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2013; 1272–1279.
149. Bregler C, Hertzmann A, Biermann H. Recovering non-rigid 3D shape from image streams. In Proceedings of the Conference on Computer Vision and Pattern Recognition, 2000; 2690–2696.
150. Xiao J, Kanade T. Uncalibrated perspective reconstruction of deformable structures. In Proceedings of the International Conference on Computer Vision, 2005; 1075–1082.
151. Del Bue A, Lladò X, Agapito L. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In Proceedings of Computer Vision and Pattern Recognition **1**, 2006; 1191–1198.
152. Wang G, Wu QM. Stratification approach for 3D Euclidean reconstruction of non-rigid objects from uncalibrated image sequences. *IEEE Trans Syst Man Cyber B* 2008; **38**: 90–101.
153. Hartley R, Vidal R. Perspective, non-rigid shape and motion recovery. In Proceedings of the European Conference on Computer Vision, 2008; 276–289.

154. Paladini M, Bartoli A, de Agapito L. Sequential non-rigid structure-from-motion with the 3D-implicit low-rank shape model. In Proceedings of the European Conference on Computer Vision, 2010; 15–28.
155. Del Bue A, Agapito L. Non-rigid 3D shape recovery using stereo factorization. *In Asian Conference on Computer Vision* 2004; **1**: 25–30.
156. Bartoli A. Estimating the pose of a 3D sensor in a non-rigid environment. In *Dynamical Vision*. Series: Lecture Notes in Computer Science, Vol. **4358**. Springer: Berlin, Heidelburg, 2007; 243–256.
157. Lladò X, Bue AD, Oliver A, *et al.* Reconstruction of non-rigid 3D shapes from stereo-motion. *Pattern Recognition Lett* 2011; **32**: 1020–1028.
158. Turk M, Pentland A. Eigenfaces for recognition. *J Cogn Neurosci* 1991; **3**: 71–86.
159. Blake A, Isard M, Reynard D. Learning to track the visual motion of contours. *Artif Intell* 1995; **78**: 101–134.
160. Tomasi C, Kanade T. Shape and motion from image streams under orthography: a factorization method. *Int J Comput Vis* 1992; **9**: 137–154.
161. Garg R, Roussos A, de Agapito L. A variational approach to video registration with subspace constraints. *Int J Comput Vis* 2013; **104**: 286–314.
162. Bue AD, de Agapito L. Non-rigid stereo factorization. *Int J Comput Vis* 2006; **66**: 193–207.
163. Vidal R, Abretske D. Nonrigid shape and motion from multiple perspective views. In Proceedings of the European Conference on Computer Vision, 2006; 205–218.
164. Chang PL, Handa A, Davison A, *et al.* Robust real-time visual odometry for stereo endoscopy using dense quadrifocal tracking. In *Information Processing in Computer-Assisted Interventions*. Series: Lecture Notes in Computer Science, Vol. **8498**. Springer: Berlin, Heidelberg, 2014; 11–20.
165. Collins T, Compte B, Bartoli A. Deformable shape-from-motion in laparoscopy using a rigid sliding window. Medical Image Understanding Analysis, 2011.
166. Hu M, Penney GP, Rueckert D, *et al*. Non-rigid reconstruction of the beating heart surface for minimally invasive cardiac surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2009; 34–42.
167. Mountney P, Yang GZ. Motion compensated slam for image-guided surgery. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI), 2010; 496–504.
168. Lourenco M, Stoyanov D, Barreto J. Visual odometry in stereo endoscopy by using PEARL to handle partial scene deformation. In *Augmented Environments for Computer-Assisted Interventions*. Series: Lecture Notes in Computer Science, Vol. **8678**. Springer: Berlin, Heidelberg, 2014; 33–40.
169. Brox T, Bruhn A, Papenberg N, *et al*. High accuracy optical flow estimation based on a theory for warping. In Proceedings of the European Conference on Computer Vision, 2004; 25–36.
170. Giannarou S, Zhang Z, Yang GZ. Deformable structure from motion by fusing visual and inertial measurement data. In Intelligent Robots and Systems, 2012; 4816–4821.
171. Wang CC, Thorpe C. Simultaneous localization and mapping with detection and tracking of moving objects. In Proceedings of the IEEE International Conference on Robotics and Automation, 2002; 842–849.
172. Wang CC, Thorpe C, Thrun S, *et al.* Simultaneous localization, mapping and moving object tracking. *Int J Robot Res* 2007; **26**: 889–916.
173. Lin KH, Wang CC. Stereo-based simultaneous localization, mapping and moving object tracking. In Proceedings of the International Conference on Intelligent Robots and Systems, 2010; 3975–3980.
174. Zou D, Tan P. Coslam: Collaborative visual slam in dynamic environments. *IEEE Trans Pattern Anal Mach Intell* 2013; **35**: 354–366.
175. Voros S, Long JA, Cinquin P. Automatic localization of laparoscopic instruments for the visual servoing of an endoscopic camera holder. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, Vol. **4190**. Springer: Berlin, Heidelberg, 2006; 535–542.
176. Voros S, Long JA, Cinquin P. Automatic detection of instruments in laparoscopic images: a first step towards high-level command of robotic endoscopic holders. *Int J Robot Res* 2007; **26**: 1173–1190.
177. Wengert C, Bossard L, Hberling A, *et al.* Endoscopic navigation for minimally invasive suturing. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, Vol. **4792**. Springer: Berlin, Heidelberg, 2007; 620–627.
178. Jayarathne U, McLeod A, Peters T, *et al.* Robust intraoperative US probe tracking using a monocular endoscopic camera. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI)*, Vol. **8151**. Springer: Berlin, Heidelberg, 2013; 363–370.

DOI: 10.1002/rcs