Abstract: Computer vision is an important cornerstone for the foundation of many modern technologies. The development of modern computer-aided-surgery, especially in the context of surgical navigation for minimally invasive surgery, is one example. Surgical navigation provides the necessary spatial information in computer-aided-surgery. Amongst the various forms of perception, vision-based sensing has been proposed as a promising candidate for tracking and localisation application largely due to its ability to provide timely intra-operative feedback and contactless sensing. The motivation for vision-based sensing in surgical navigation stems from many factors, including the challenges faced by other forms of navigation systems. A common surgical navigation system performs tracking of surgical tools with external tracking systems, which may suffer from both technical and usability issues. Vision-based tracking offers a relatively streamlined framework compared to those approaches implemented with external tracking systems. This review study aims to discuss contemporary research and development in vision-based sensing for surgical navigation. The selected review materials are expected to provide a comprehensive appreciation of state-of-the-art technology and technical issues enabling holistic discussions of the challenges and knowledge gaps in contemporary development. Original views on the significance and development prospect of vision-based sensing in surgical navigation are presented.

1 Introduction

1.1 Background

Surgical navigation in computer-aided surgery addresses the many challenges faced in modern surgery. Minimally invasive surgery (MIS) is an example of an emerging standard adopted by surgeons. While MIS reduces trauma on a patient through small incisions and exposing them to lower risk of medical complications, such procedures are often associated with challenging visual, perceptual, and dexterous constraints. Computer-aided surgery addresses these issues by facilitating presurgical diagnostic imaging and planning [1–4], intra-operative visual guidance [5–8], and surgical robotic applications [9–13]. The challenges and solutions are summarised in Fig. 1 with the classified references in their appropriate categories.

An important provision to the above-mentioned capabilities is spatial information processing of the operation field. This includes tracking and localisation through position sensors data acquisition, digital image processing, machine vision, or a combination of these methods. Amongst those relevant fields, the vision-based technique has contributed extensively in Surgical Navigation technology to support effective computer-aided MIS. It offers real-time acquisition and localisation of surgical site for image overlay onto pre-operative medical images and anatomical models. Vision-based metrics also facilitate robotic control and augmented reality implementation in computer-aided surgery.

1.2 Other review papers

There have been numerous prominent review papers that discuss the progress in surgical navigation [5, 6, 10, 14–18], including a few related to the application of computer vision [5, 6, 10, 17]. Stoyanov [6] presents one of the most comprehensive reviews on surgical vision focusing on the aspect of enhanced biophotonic imaging. It also covers concepts in quantitative endoscopic vision and tissue morphology tracking, though with limited details in these aspects compared to the review by Mountney et al. [17]. A more application-focused survey on vision-based navigation in image-guided interventions is published by Mirota et al. [5]. It covers quantitative discussion on state-of-the-art technologies and existing systems deployed in the various field of medical interventions like Rhinoscopic neurosurgery, laparoscopic surgery robot-assisted surgery and Natural Orifice Transluminal Endoscopic Surgery (NOTES). A more recent literature review by Bouget et al. [19] on vision-based surgical tool detection and tracking provides an analytical coverage of the detection methods together with the data set and the techniques for validation. Readers may be interested in the book chapter by Speidel et al. [20] for an up-to-date introduction to the subject of vision-based intervention imaging. The chapter encompasses vision-based intervention imaging-related topics from imaging modalities to scene analysis, to scene interpretation and, finally, its clinical applications.

1.3 Scope and organisation

Although there are substantial published reviews on surgical navigation, specific reviews on vision-based techniques in surgical navigation in the context of computer-aided surgery are limited. Holistic reviews [5, 6, 18, 19] in this technology are only available recently considering the beginning of robotics surgery and computer-aided surgery in the early nineties. There is yet to be a consensus on the representative taxonomy in the classification of approaches and developmental milestones in this field over these years. Hence, this review will not attempt to present an exhaustive generalisation of the subject. Instead, this review aims to cover a comprehensive survey on selective materials and technical discussion on the vision-based techniques in contemporary works followed by the authors’ view on this subject. The scope will be organised in a self-contained manner described as follows.
A survey of current literature that contributed to the foundation of classical theories pertaining to computer vision in computer-aided surgery will be presented in Section 2. This will be followed by a thorough review and analysis of contemporary researches in the area of vision-based technique for computer-aided minimally invasive surgery in Section 3. Section 4 discusses the state-of-the-art technologies and applications shaping modern surgical navigation approach in minimally invasive surgery. Representative articles will be selected for a case study where the in-depth technical discussion will be presented in Sections 3 and 4. Through the above-mentioned series of rigorous reviews, the authors will attempt to present their original views to the challenges, impacts, and future developments of vision-based tracking for surgical navigation in Section 5.

2 Principle of vision-based tracking
The theoretical basis of vision-based tracking technology stems largely from the subject of computer vision. Due to its broad association with various applications and technologies, reviews papers in computer vision references usually present their survey in a unique perspective relevant to their respective application. It does not have much significance, nor is it the intention of this review to present an exhaustively unbiased survey of the broad nature of vision-based tracking. However, this section examines the principles and research topics that shaped the theoretical framework of vision-based tracking relevant to surgical navigation. The intention is to establish a common and relevant background for the specific discussion of contemporary researches and applications in the field of computer-aided surgery and surgical navigation.

2.1 Camera pose estimation and multiple view geometry
A key concept in vision-based tracking is the unification of camera geometry and multiple view geometry. There exist immense references [21–23] that introduce the principle of the camera and multiple view geometry. Camera geometry, in general, aims to establish a camera calibration matrix that relates world coordinates to image coordinates based on an assumed physical model of the camera while multiple view geometry involves the computation of a transformation relationship between camera views. By fusing these two concepts, obtaining camera motion and the surrounding scene is made possible. To illustrate, assume a pinhole camera projection model representing the mapping of 3D points $M = [X Y Z]^T$ onto points $m = [x y 1]^T$ on the image plane as shown in Fig. 1a expressed by

\[ m = K[R \ | \ t]M = PM, \]  

where $K$ is the intrinsic matrix representing the camera's innate properties; $R$ and $t$ represent the rotation and translation of the camera also known as extrinsic parameters. The product of the intrinsic and extrinsic matrix is known as the calibration matrix, $P$.

Camera pose estimation is often essential for the application of vision-based surgical navigation. In the case of a monocular camera, the problem of interest is to estimate the position and orientation of the camera given a set of correspondence between some interesting points in the 3D space and their image coordinates in that camera views. Mathematically, this is a problem of solving for the projection matrix $P$ in (1) given a set of 2D–3D point correspondences $(x \mapsto X)$.

The direct linear transformation (DLT) [24] can be used to solve the 11 parameters in $P$ without considering prior knowledge of intrinsic parameters. Each correspondence $(x_i \mapsto X_i)$ can be used to represent two linear equations

\[ x_i = P(1,:) \cdot X_i/P(3,:) \cdot X_i \]  

\[ y_i = P(2,:) \cdot X_i/P(3,:) \cdot X_i \]

where $(x_i, y_i)^T = x_i$ and $P(m,:)$ implies all row $m$ entries of $P$. This set of the linear equation can be written as the ‘$Ah = 0$’ problem ignoring the trivial solution of $h = 0$. Singular value decomposition (SVD) of $A$ can be used to obtain the solution.

However, this approach usually requires a large number of interest points, which is unlikely in tissue imaging. There is no guarantee that the desirable geometrical distribution of the point correspondences can be achieved since the corresponding points are passively extracted features. Hence there is a high risk that DLT, which relies on over-parameterisation, may be ill-posed under the unfavourable correspondence layout.

The fact that the intrinsic parameters are innate camera properties not related to the camera pose, it is possible (in fact, more widely recommended in the literature [25, 26] for 3D tracking) to be estimated separately through a camera calibration process under a more controlled set up. Camera calibration can be done with a known structure like a checkerboard pattern, as shown in Fig. 2. The pose can be recovered through the corresponding known interest points between multiple views to optimise on the set of intrinsic parameters for the camera, as shown in Fig. 2.

Since the information of the camera's intrinsic parameters is readily available and can be verified across independent calibration processes for consistency, this useful prior knowledge should be
utilised. Therefore, it is more appropriate to formulate the endoscopic pose estimation as a perspective-n-points (PnP) problem [27–29] solving the extrinsic matrix with prior knowledge of the intrinsic matrix.

The general framework for camera pose determination includes solving the calibration matrix, which comprises of the intrinsic and extrinsic matrix. From the epipolar geometry depicted in Fig. 1b, correspondence condition between points $x$ and $x'$ in an image plane of two different perspectives can be mathematically stated in terms of the fundamental matrix, $F$ as in

$$x^T F x = 0.$$  \hspace{1cm} (4)

In general, it can be shown [21] that

$$F = [e^e]^\top P P^\star,$$  \hspace{1cm} (5)

where $[e^e]$ is the $3 \times 3$ skew symmetrical matrix of the epipole, $e^e$ and $P^\star$ is the pseudo inverse of $P$ (Fig. 3).

### 2.2 Stereovision

Another practical implementation approach is the use of stereovision [30, 31]. This is a special case of multiple camera view using the epipolar constraint. The concept is to use image disparity as a to recover depth information. This subject has great influence in shaping the developmental scene of vision-based tracking. It is of particular importance to surgical navigation because of the promising development in 3D display and stereoscopic endoscope for the minimally invasive procedure [32]. Together with many technological factors, it is responsible for the emergence of stereoscopic endoscope. Many of the works [31, 33–37] that will be discussed in the latter part of this paper involve the application of stereovision for tracking and localisation.

Stereo vision based applications generally involve four steps, namely, lens undistortion, stereo rectification, correspondence and reprojection through triangulation. Calibration of a stereo rig usually includes the first two procedures. The general idea is to rectify the distortion of the lens and relative position of the pair of cameras to facilitate triangulation. Lens distortion is compensated by radial and tangential distortion function. This rectification step is done during the calibration process, as mentioned previously (Fig. 2). The coordinates of the pair of camera plane are rectified such that they are row-aligned. A calibrated and rectified stereoscopic endoscope can locate the position of a point in 3D space through triangulation [38].

### 2.3 Feature tracking

Detection of interest points is an important prerequisite to computer vision application. It can be used for automatic localisation of corresponding points between multiple images for the computation of spatial information. Schmid et al. [39] defined interest point as any point within an image that changes signal two-dimensionally. Detection of interest point is basically obtaining visual cue from an acquired image. In the case of surgical navigation, this is essential to perform image registration, surface reconstruction, and 3D localisation. Features on the 2D image need to be detected, described and matched with their correspondence in the subsequent frame before we can apply camera geometry to establish a spatial relationship between the camera and world coordinates, as discussed in the previous two sections. An active approach requires a fiducial marker or artificial landmarks to be included in the scene. The model-based approach and nature feature detection approach do not require the specification of the imaging environment. In fact, the latter has been a more viable
approach for both clinical and research implementations. The steps from camera calibration to registration of multiple views are illustrated in Fig. 4 with an example of underwater camera especially in the surgical environment due to dynamically changing performance of 21 descriptors with respect to the deformation of calibration for the application of placenta foetoscopy.

Table 1

| Category | Camera (endoscope) | Environment (surgical scene) | Reviewed articles |
|----------|--------------------|-----------------------------|------------------|
| 1        | random motion      | static                      | RA1-7            |
| 2        | random motion      | changing                    | RA8, 9           |
| 3        | fixed/controlled   | changing                    | RA10-28          |
| 4        | machine-learning   | changing                    | RA29-30          |

Both Categories 1 and 2 are interested in camera pose estimation. The main objective in Category 1 is to track the endoscope motion for medical image registration, image-guided catheterisation, or diagnostic intervention. In Category 2, the problem is complicated by the dynamically changing environment. Such applications are usually interested in both endoscope motion tracking and the recovery of moving anatomical features. The tracking problem can be better determined under a more controlled operation condition where the camera position is fixed or known (category 3). Cases associated with this category are interested in the motion of soft tissue or other surgical tools.

3.2 Anatomical structure and endoscope motion estimation

The subject of endoscope motion estimation and anatomical feature tracking has been proposed in several studies. In this section, we will discuss selected work [35, 46–51] on endoscopic vision-based tracking. These selected works adopted a common approach that exploits endoscopic surgical imaging instrument as the visual source of information and usually adopted the assumption of the camera in random motion, i.e. Category 1 or 2. Depending on its theoretical treatment of changes in the acquired scene, they either assume a static environment (Category 1) or take into account changes in the surgical scene (Category 2).

3.2.1 Medical image registration-based support: A common approach adopted in the tracking of a flexible endoscope is through registration between real endoscopic images and images from the virtual endoscopic system (VES). This is a Category 1 problem, as most cases discussed here use pre-operative imaging like computed tomography (CT) without accounting for spatial–temporal changes and deformation during surgery.

Unlike a rigid endoscope, a flexible endoscope does not have the option of direct motion estimation using an external tracker since the tip of the endoscope cannot be obtained via rigid transformation. Modelling the tip motion using boundary
conditions and dynamic constraints through external tracker is possible but belongs to a totally different and challenging research area. Fortunately, the path of the flexible endoscope is very much constrained by the luminal wall, which can be constructed from pre-operative imaging devices. The highly constrained possible endoscopic scene enables its position to be reasonably computed by referencing medical images. Hence, the image registration-based approach between virtual and real information has been a popular approach for flexible endoscope tracking. Early work for bronchoscope tracking was proposed by Mori et al. [47]. As this is one of the complete early works on VES, we will review it in detail as a technical case study.

**Case Study 1 (RA11): Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images.**

**Background and rationale:** The navigation system known as the virtual endoscope systems are widely used to display the virtual scene of the 3D anatomical model of the respiratory tract constructed from the patient's CT scan. Such a system provides surgeons with navigational information in complement to the actual endoscopic view. As the virtual scenes were rendered from pre-scanned CT data with texture mapping, it does not reflect intra-operative changes and actual surface information for clinical examination. A real endoscope is essential for clinical examination as well as a timely situation update of the anatomical profile. However, the real endoscope does not provide the user with navigational information from a global perspective. There are numerous bifurcations to navigate through before reaching a desired location of the bronchus. Hence, there is a need to register a real endoscopic scene to the virtual endoscopic scene to provide the surgeon with navigational information and, at the same time, updated intra-operative and anatomical surface information.

**Method:** In this work, epipolar geometry analysis and intensity-based image registration of the real-virtual endoscopic image were used to track and compute bronchoscope's 3D position. The camera pose estimation problem was formulated as an optimisation of the image similarity of the real endoscopic image and virtual endoscopic image. The general approach is to perform a rough estimate of the camera motion using epipolar geometry followed by intensity-based registration for precise estimation. Optical flow using simple block matching was used to obtain corresponding points with maximum cross-correlation based on a Powell search of the local descriptor. The corresponding points were subsequently used to compute the fundamental matrix using Singular Value Decomposition. A rough estimation of the extrinsic camera matrix is obtained through epipolar geometry by solving (4). Subsequently, the rough estimation is used to obtain an estimation of the position with respect to the virtual environment. Precision estimation is then computed by another round of intensity-based registration with the maximisation of the ratio of cross-relation and mean square error between real endoscopic image and virtual endoscopic image, as illustrated in Figs. 5 and 6.

**Experiment and results:** The results were promising. There were 600 consecutive frames successfully tracked in the best scenario, and the tracking was sufficiently accurate based on visual inspection. It was, however, reported that the computation time for processing a frame is about 20 s. While the robustness and accuracy appear promising, its clinical application at the time of publication may very much be limited by the computational performance.

Slow processing time of method featured in Case Study 1 was, however, addressed in another work [52] by the same author. A hybrid system was implemented where a magnetic tracking sensor was incorporated in addition to the image registration module. For large breathing motion the accuracy was about 3 mm and processing rate of 0.8 frames per second. This approach was further improved in another work [53] with the implementation of an algorithm for selective comparison of image region for effective processing. The system can track up to 1600 frames in ~50 s.

Apart from VES applications, medical image-based registration using visual tracking means is also common in intra-operative surgical navigation where a pre-operative plan for interventional procedures is mapped to intra-operative robotic execution as featured in RA23 [54]. Fig. 7 shows the workflow of the registration between the preplanned surgery and actual execution via a vision-based approach.

Image registration of multiple imaging modalities for surgical visualisation of endoscopic procedures with extremely limited field-of-view (FOV) and dexterity is another common application. One example is foetal surgery, where a fetoscope is used. This specialised non-flexible endoscope images the uterus during the minimally invasive procedure with the camera motion constrained to the incision while FOV is limited due to the miniature size of the scope. In RA24, image mapping between the fetoscopic view and intra-operative ultrasound [55, 56] has been explored to facilitate visualisation in surgical navigation. The fusion of the two intra-operative imaging modalities is illustrated in Fig. 8. This 3D visualisation platform of the expanded photorealistic endoscopic view on the 3D ultrasound constructed model provides enhanced visual augmentation for surgical navigation during fetoscopic laser photo-coagulation treatment.

**3.2.2 SLAM framework:** Development in robotic navigation technology also influences vision-based tracking in surgical navigation. Simultaneous localisation and mapping (SLAM) has been a common algorithmic framework for surgical navigation and vision-based associated techniques are especially prominent in the adoption of this framework.

One of the earliest examples that extended the SLAM framework for medical image processing was for registration of monocular endoscopic images to CT scans proposed by Burschka et al. [57] for sinus surgery. This method requires pre-operative imaging data from CT scan and accurate reconstruction of the anatomical model. This may not always be feasible, depending on the anatomical case. While technology in medical image registration has been well established as we have seen in previous studies, there remain limitations in accuracy due to geometric misalignment between pre-operative and intra-operative imaging. Anatomical features may appear different due to different biological or mechanical conditions during diagnostic imaging and surgery. A ‘sequential vision only’ approach was presented by Mountney et
al. [46]. This is another early example in formalising endoscopic image-guided surgical navigation with the SLAM framework. The problem statement is defined as the simultaneous localisation of stereoscopic soft tissue mapping for minimally invasive surgery. This is in alignment with the SLAM framework that estimates moving sensor while building a reconstruction of the observed scene. The method relies solely on sequential visual acquisition to compute 6 DOF camera motion. The feasibility of using a stereoscopic fibroscope to recover camera motion and 3D scene information was studied by Noonan et al. [58]. In this work, image transmission from distal fibroscope tip to proximally mounted CCD camera was designed to concurrently track stereo imaging instruments and mapping of the soft tissue surface. A sequential vision, only the EKF-SLAM approach, was adopted.

In assuming smooth camera motion, constant velocity and acceleration constraints are used to derive the motion model. This is, however, not a realistic assumption in most cases of hand-held endoscopic application and may result in system failure. The EKF SLAM might have to be complemented by other algorithmic procedures to address the relaxation of the smooth motion assumption. Grasa et al. [51] exploited the processing efficacy of EKF monocular SLAM to construct a photorealistic 3D model, spatial measure of anatomical feature and augmented reality (AR) applications via the freehand motion of a standard hand-held endoscope. In this work, the authors focused on an EKF + ID + JCBB monocular SLAM approach. This method was claimed to be the state-of-the-art approach for medical images by the author at the time of publication. The general performance is said to be enhanced with the inverse depth (ID) for map point coding due to the improvement in the linearity of the measurement equation. To enforce scene rigidity, Joint Compatibility Branch and Bound (JCBB) is used to reject spurious points. This study is based on the assumption of the rigid environment (intervention cavity) and non-pure rotation motion, which, according to the authors, are common conditions for many laparoscopic surgeries, for instance ventral hernia repair.

However, most surgical applications involve severe soft tissue deformation due to non-deterministic tool–tissue interaction. Moreover, the endoscope moves in and out of the abdomen frequently in practice. Hence, Grasa et al. [48] further customised the EKF monocular SLAM to enhance the robustness of laparoscopic localisation and tracking. Unlike the EKF + ID + JCBB framework proposed in [51], this work adopts an EKF + 1-point RANSAC [59] + randomised list recalocalisation to cope with sudden camera motions or laparoscope reinsertion.

A major challenge in vision-based tracking associated with the anatomical structure is tissue deformation. Despite an adopting enhanced SLAM framework, this problem continues to be an issue. Traditionally, the SLAM theoretical approach to non-static motion is to treat them as outliers. This may not be desirable in the dynamic environment of MIS, especially during moving heart surgery. The below case study proposed a motion compensation model for the SLAM framework.

**Case Study II [RA9]: Motion-compensated SLAM for image-guided surgery**

**Background and rationale:** Registration of pre-operative and intra-operative information for surgical navigation can be challenging due to the large tissue deformation associated with cardiac, gastrointestinal, or abdominal surgery. To manage deformation at a reasonable meaningful extent, Mountney et al. [60] proposed the incorporation of rhythmic respiratory and cyclical cardiac motion models into the SLAM framework to manage the dynamic scene. This approach was termed motion compensation simultaneous localisation and mapping (MC-SLAM).

**Method:** An online deformation learning method [61] to track regions of interest is adopted for robust tracking of the tissue surface. The spatial–temporal information of the tracked feature points in 3D motion coordinates are subsequently correlated to respiratory motion using principal component analysis (PCA). Output data from PCA is curve fitted to a typical function for respiratory cycle [62] represented by

\[
\phi(t) = z_o - b \cos^n \left( \frac{\pi t}{\tau} - \phi \right),
\]

where \( z_o \) is the position of the liver exhale, \( b \) is the amplitude, \( \tau \) is the respiratory frequency, \( \phi \) represents the phase and \( n \) describes the shape of the model. Levenberg–Marquardt minimisation algorithm is used to optimise the parameters.
The MC-SLAM framework defines the state vector with an additional component, \( \dot{m} \). The state vector and covariance are represented as

\[
\dot{x} = \begin{bmatrix} \dot{x}_v \\ \dot{m} \\ \dot{y} \end{bmatrix}, \quad \text{where} \quad \dot{x}_v = (r, R, v, w)^T, \quad \dot{m}
\]

\[
= (\alpha, \tau, b, z_w, \gamma) \quad \text{and} \quad \dot{y} = (\dot{y}_1, \ldots, \dot{y}_n)
\]

Variables \( r, R, v, \) and \( w \) represent the position, orientation, linear velocity and angular velocity, respectively. Elements of \( \dot{m} \) are the parameters in equation with \( \alpha \) being \( \pi/\tau \). The notation \( \dot{y}_i \) is the \( i \)th map feature derived from PCA. The modified measurement model is expressed as

\[
h^R_i = R^{RW}(\varepsilon_{z_o} - b \cos^{2n}(\alpha) + \dot{y} - r^R)
\]  \( (7) \)

Fig. 9 illustrates the modified SLAM framework.

**Results:** Evaluation of tracking performance based on simulated data produced a mean error of 0.25 and 1.31 cm for MC-SLAM and static SLAM, respectively. Camera localisation accuracy in ex vivo study registered a mean error of 0.11 cm for MC-SLAM and 0.56 cm for static SLAM. The application was also demonstrated in vivo conditions. Ground truth data was, however, not available for accuracy evaluation during in vivo study.

### 3.3 Soft tissue tracking

Over the past decade, there has been an active research effort in the introduction of a vision-based approach for tracking of tissue morphology and motion recovery. This is due to both technological push factors associated with the development of a more capable computer vision system and pull factors motivated by the urgent need for intra-operative feedback of spatial-temporal information in soft tissue deformation. Modern surgery where robot-assisted surgery and sophisticated image-guided visualisation system are increasingly used, it is important to acquire information on the dynamic changes due to many factors like cardiac, breathing, or random tool-tissue interaction. The application of vision-based tracking for motion recovery of tissue surface is especially prominent in the human–computer interface of robotics and image-guided navigation. The design goal for this application is to virtually stabilise moving surgical scene. Many of these works revolve around the application in cardiac surgery for the tracking of heart surface. It appears that a vision-based approach provides one of the most streamlined mechanisms in the case of the minimally invasive procedure, as there is no need for additional sensors. This is a very important advantage as it does not strain on the demanding workspace limitation in such a procedure.

One of the first vision-based systems for the mentioned application was proposed by Nakamura et al. [63]. This is a representative work of active approaches in vision-based technique in surgical navigation and will be discussed in detail in the following case study.

**Case Study III [RA110]: Heartbeat synchronisation for robotic cardiac surgery**

**Background and motivation:** In off-pump heart minimally invasive direct coronary artery bypass (MIDCAB), surgeons operate on beating heart. The surface motion of the heart makes the operation challenging and existing mechanical heart stabilisers do not provide surgeons with complete stabilisation. There remains residual motion. This study aims to propose a surgical robot capable of synchronising with the moving heart during MIDCAB to achieve motion compensation. Three technologies were developed in this work, namely visual synchronisation, motion compensation, and master–slave control. We will discuss the vision-based technology and its integration with other components of the system in this section.

**Method:** The system comprises a visual interface and a master–slave robot. It tracks fiducial markers attached to the heart surface with a high speed camera to achieve visual and motion stabilisation against cardiac and respiratory motion through visual feedback with a high-speed acquisition rate of 995 fps. Also, a heartbeat prediction model using the \( p \)th autoregressive model was incorporated to supplement the heart surface motion recovery. Together with the motion detection from the high-speed camera, a colour video in NTSC format is processed to display the scene of a visually stable heart surface while the slave manipulator arm executes motion-compensated trajectory output via the master–slave control system. Transformation of command in master input-domain to slave task-domain is achieved through differential kinematics and motion compensated through visual servo. The framework of this system is illustrated by Figs. 10a–d.

**Results:** The system was evaluated on an experiment where a tool is held by the slave manipulator to move towards an oscillating laser point controlled by the user from the master manipulator. In this experiment, the visual interface displays the tool moving steadily towards a stationary laser point. The registered deviation exhibited by the slave manipulator from the combined signal of master and reference is <0.5 mm. Further validation was carried out on in vivo experiment, as shown in Fig. 5e.

**Comment:** While the work appears promising, there was no detail on the tracking algorithm and the fiducial marker placement strategy. Despite the promising potential, the use of a high-speed camera may not be readily feasible in clinical practice. A less technology-dependent solution is to address the underlying issues.
in information processing through the effective implementation of an algorithm. In addition, the need to fix passive markers on anatomical structure via limited tool entry points poses implementation challenges for actual medical use. Moreover, the system is limited to surgery with a relatively localised surgical site. General abdominal surgery may involve a wider surgical site covering various anatomical features instead of a limited surface region.

Other active approaches include a structured light system to recover shape [64] and laser plane sweep over the surgical site [65]. These methods are not popular in the MIS application as they require an additional incision for instrument entry. A passive vision-based tracking approach may be more appropriate to extend the range of applications. The passive approach relies on the natural scene and do not require the introduction of artificial markers. Several methods adopted by researchers to enhance the reliability of the passive approach include establishing a probabilistic tracking framework, defining geometrical constraints, designing a machine learning algorithm, applying a motion predictive model, and approximation of surface deformation.

Probabilistic framework: Lo et al. [33] proposed a probabilistic approach that uses Markov random field-based Bayesian belief propagation framework to fuse multiple depth cues for deformable 3D surface reconstruction. The depth cues are obtained from surface shading-based reconstruction techniques and feature-based stereoscopic correspondence. In addition, a belief propagation scheme that utilises sparse stereo points for depth inference of surrounding path is proposed in the study. Another probabilistic framework that fuses tracking information based on different descriptors is also proposed by Mountney et al. [41].

Geometrical constraint: Stoyanov et al. [66] combined sparse salient feature tracking with the surface model to track tissue deformation for cardiac surgery. However, the tracking is only maintained within the first 5 s. In a later work, Stoyanov and Guang-Zhong [36] proposed the incorporation of surface geometrical constraints to further enhance the motion surface tracking.

Computational intelligence: The approaches discussed thus far relied mainly on 'ad hoc' representation as termed by Mountney and Yang [61]. According to them, such an approach is inadequate for representing the dynamic motion of the soft tissue deformation. The proposed online learning of the local deformation approach to the problem of soft tissue motion recovery hence achieving robust feature tracking. In their study, it was argued that the sole reliance of vision technique from the natural scene is inadequate suited for MIS due to issues like inter-reflection variation and instrument occlusion. Hence, an online learning feature-based tracking without restriction to any specific image transformation and visual characteristic was proposed. The proposed approach was benchmarked against existing algorithms and validated on both simulated and in vivo cardiovascular and abdominal MIS data. Motion source separation using ICA to decompose the cardiac and respiratory motion was also demonstrated. Automatic instrument segmentation using the deep learning approach is by Shvets et al. in RA30 [67]. The use of deep learning has not been as extensive in the application of surgical navigation compared to other medical or industrial domains.

Predictive model: Apart from visual acquisition, the incorporation of a motion prediction model is also proposed. Richa et al. [49] applied a predictive extended Kalman filter based on a time-variant dual Fourier series. This method exploits the quasi-periodic heart motion to predict the upcoming motion. Like the work by Grasa et al. [48, 51], it aims to enhance tracking failure recovery due to occlusion and other motion disturbance.

Surface deformation model: Deformation representation can also be adopted to track surface motion and deformation. In such cases, parameters of local time-varying algebraic descriptions are automatically tuned for optimal inference of the 3D structure. Lau et al. proposed the use of B-spline for surface representation, justifying his choice based on the size of parameters required, continuity, uniqueness, local shape controllability, and linearity. Richa et al. [34, 37, 68] further proposed a vision-based motion compensation method by incorporation of thin-plate spline (TPS) deformation model. This theoretical concept will be discussed in detail under the subsequent case study based on [37].

Case Study IV [RA21]: 3D motion tracking for beating heart surgery using a thin-plate spline deformable model

Background and rationale: Like many other types of research work in motion compensation and soft tissue tracking, this work suggests that a virtual stable operating field can be achieved through synchronisation of tasks with the recovered motion of the organ surface. Two major challenges for such applications identified in this article include the complex dynamics and poor illumination condition of the surgical scene. Hence, the TPS parametric model is extended for vision-based 3D tracking of the complex surface region of a moving heart and an illumination compensation procedure design for beating heart conditions.

Method: The TPS formulation extended to model the heart surface depth with points of 3 degree of freedom is represented in (8). The projection of control points onto the image can be represented by the mapping, \( m(\mathbf{x}) \), as illustrated in Fig. 11.

\[
m(\mathbf{x}) = \begin{bmatrix} r_1^x & r_2^x & r_3^x & r_1^y & r_2^y & r_3^y & r_1^z & r_2^z & r_3^z \end{bmatrix}^T + \sum_{i=1}^{w_0} \begin{bmatrix} w_{0i} \end{bmatrix} U( \parallel \mathbf{c}_i - \mathbf{x} \parallel ) \tag{8}
\]

The mapping can be written in the following form:

\[
P = MK [\tilde{P} \tilde{z}] \tag{9}
\]

The expression used in the article requires rigorous step-by-step derivation beyond the capacity of this case study discussion. To preserve the consistency of the expression without over reproduction of details, (10) can be taken as a transformation that maps \( [\tilde{P} \tilde{z}] \) to \( P \), where \( P \) is the pixel coordinates that undergone the warping; \( \tilde{P} \) and \( \tilde{z} \) are the stacked coordinates of control point and depth position, respectively. By recognising that the positions of the control point correspondences are in fact 3D point projection onto the image plane, the stacked coordinates of the control point \( [\tilde{P} \tilde{z}] \) can be expressed in terms of the camera calibration matrix, \( C \) as follows:

\[
[\tilde{P} \tilde{z}] = \begin{bmatrix} \mathbf{X}^T \mathbf{1} \end{bmatrix} = [\mathbf{X} \mathbf{1}] C^T \tag{10}
\]

Hence (10) can be rewritten as

\[
P = MK[X \ 1] C^T \tag{11}
\]

The 3D warping function in terms of the pixel coordinates, point coordinates and camera matrix is expressed formally as

\[
w_{3D}(x_i, h, C) = MK_h[h^s \ h^r \ h^l \ 1]^T \tag{12}
\]

where \( MK_h \) can be seen as the ith term of \( MK \), associated with \( x_i \).

With a calibrated 3D stereo rig, we can estimate the 3D coordinates of points that map to image plane minimising the alignment error between a reference image and the left-right stereo image concurrently.

The problem is formulated as follows:

\[
\min \epsilon = \sum_{h} \left[ I_l(w_{3D}(x_i, h, C)) - T(x) \right]^2 + \left[ I_r(w_{3D}(x_i, h, C)) - T(x) \right]^2 \tag{13}
\]

where \( A \) is the set of template coordinates, \( I_l(w_{3D}(x_i, h, C)) \) is the transformed image, subcript \( l \) and \( r \) represent left and right, respectively, and \( T \) is the reference image. The efficient second-order minimisation (ESM) algorithm based on iterative registration for tracking was adopted to solve the minimisation problem [69].
Illumination compensation is also introduced to deal with large deformation and poorly illuminated environment. This problem is formulated as finding the element-wise multiplicative lighting variation for each pixel $x_i$ of the current image $I$ and a global bias $b$ such that $I$ coincides with the closest illumination conditions of $T$.

$$I'(x_i) = I(x_i)I(x_i) + b$$  \hspace{1cm} (14)

The same optimisation framework was used to estimate the illumination compensation parameters

$$
\begin{align*}
\min_{h, y, \delta, y'} & \quad c = \sum \left[ (I'(x_i)T(x_i) - I(x_i))^2 + (I'(x_i)T(x_i) - I(x_i))^2 \right] \\
\text{s.t.} & \quad \sum \left[ (I'(x_i)T(x_i) - I(x_i))^2 + (I'(x_i)T(x_i) - I(x_i))^2 \right] 
\end{align*}
$$  \hspace{1cm} (15)

Specular highlights were detected and removed from the image before the estimation of illumination and warping function. This was done by constructing a secularity map through thresholding of intensity. Dilation using a circular kernel was subsequently carried out to enhance the map.

**Outcome:** In an ex vivo study based on a 4 cm x 4 cm region of 25 control points, the average error is 0.24 mm. Tracking can run at a video frame rate of 50 Hz with a computation time of 18 ms within frame intervals.

### 4 State-of-the-art

Although vision-based tracking in computer-aided surgery was introduced fairly recently, it has shown promising potential for surgical navigation application in various aspects of surgery and medical intervention. In addition, this approach plays a contributive role in the development of potentially impactful research areas like robotic surgery, and augmented reality surgical system. In this section, we review the level of technological development for this approach in different domains of medical intervention. Its contribution to modern surgery, including representative technologies like robot-assisted surgery and augmented reality surgical guidance, will also be discussed based on selected articles.

#### 4.1 Surgical application

While many of the discussed developments are still in their early research phase, surgical navigation via vision-based tracking has certainly permeated into the various fields of surgery and medical intervention. In the inspection of the respiratory tract, the common approach of image registration is usually adopted, as seen in [47, 52, 53]. This is partly due to the nature of the application, which essentially represents a problem of navigating a flexible endoscope in a semi-rigid constrained path of the lumen. Traditionally, medical images acquired prior to the interventional procedure were used as a guide for navigation. This approach requires the practitioner to mentally register pre-operative information, including a surgical plan to intra-operative endoscopic imaging. Anatomical changes from the time of imaging to the actual procedure may result in confusion and error. Studies like [47, 52, 53] aim to address these issues. There has been a substantial improvement in processing performance over the past decade. Early work by Mori et al. [47] reported a processing rate of about 20 s for a frame in 2002. However, recent works by Mori et al. [52] and Deguchi et al. [53] documented processing performance of 0.8 and 0.03 s, respectively. A similar image registration approach is also commonly used for the inspection of the skull via a nasal cavity and neurosurgery, as featured in [57].

State-of-the-art vision-based surgical navigation is also well represented by surgical applications in abdominal and cardiac–thoracic operation. Problems in these domains are complicated by the dynamically changing scene and relatively unconstrained endoscope motion in free space. Studies discussed thus far appear promising in their problem-specific studies. However, very few address the free form of camera motion and dynamically changing environment holistically. The work by Mountney et al. [60] appears to be one of the complete works in this aspect, addressing both free-form camera motion and anatomical motion. However, in vivo evaluation of the motion model remains intractable due to the unavailability of ground truth. Nevertheless, the incorporation of its motion model produces superior tracking performance in comparison to the convention model that assumes a static environment.

Table 2 represents the classification of the discussed research and development work in their various related surgical application domains.
4.2 Robot-assisted surgery

Vision-based tracking in navigation for robot-assisted surgery constitutes a significance representation of state-of-the-art technology in surgical navigation. In fact, the contribution of surgical navigation is only meaningful when it provides effective intra-operative guidance that translates into precision and accuracy for clinical application. Visual guidance alone does not ensure consistency in the execution of surgical tasks especially under demanding operation conditions like those of minimally invasive surgery. Robotics is an excellent candidate for managing precision, and accuracy has been widely developed and researched to bridge this knowledge gap. Today, robot-assisted surgery can be performed by a commercially available system. Many of the discussed vision-based technologies were directed at facilitating robot-assisted surgery [36, 50, 63, 66, 70].

An important technology in this aspect is motion compensation through vision-based feedback for virtual stabilisation. Motion synchronisation as termed by Nakamura et al. [63], can be achieved through visual servoing, as discussed previously. While this example demonstrated the feasibility of visual and motion synchronisation on a customised robotic arm, robot-assisted surgery integrated with vision-based surgical navigation has been shown to work fine with commercially available platforms, like da Vinci robotic surgical system (Intuitive Surgery Inc.), or AESOP manipulator (Computer Motion Inc.) in [36, 50, 66, 65] respectively. The state-of-the-art surgical robot navigation can be represented by clinically useful features like safety management. The following case study demonstrates the application of an active vision-based approach for surface reconstruction and safety management in surgical navigation.

Case Study V [RA11]: Laser-scan endoscope system for intra-operative geometry acquisition and surgical robot safety management

Background and motivation: The paper stated the relevance and significance of the surgical robot in modern surgery. It is believed that intra-operative organ structure and information are crucial for surgical robot navigation, including features like obstacle avoidance and safety management system that avoid undesirable injury.

Approach: A laser-scan endoscope was developed to acquire laparoscopic real-time 3D visualisation of surgical objects with video-texture mapping. Shape reconstruction and 3D point measurements were performed during laparoscopy. Robotic navigation and safety management was demonstrated through inter-sensor registration.

Method: The system configuration consists of a high-speed endoscope camera, a laser emitting endoscope, and an infrared external tracking system that acquires the position of camera and laser device. A laser-beam strip is controlled by an optical galvano scanner while the high-speed camera detects the laser beam line to obtain the 3D geometry of a given organ. A task parallel processing architecture was implemented to manage acquisition and visualisation. Two PCs were used to execute the measurement task and visualisation task concurrently. The scanned data was stacked into shared memory located in one of the PC while the other retrieves and constructs the surface. The scene provides 5–6 fps frame rate visualisation of the organ deformation updated using OpenGL.

The procedural outline is illustrated in Fig. 12. Controlled laser projection onto the organ’s surface is captured by a high-speed camera to compute 3D point measurement and reconstruct surface profile. Texture mapping is then carried out to create a photorealistic model that is incorporated into the virtual navigation system. User interaction with the robotic system and surgical site is displayed intuitively on the virtual interface and executed by a robot arm in reality.

Contribution: The topic of interest in this case study is, however, the safety management feature in its robotic navigation system. With the developed geometry-based navigation system, the motion of the surgical robot and its interaction with the surgical site can be readily monitored by the user and the system built-in feature. The surgeon can judge intra-operatively spatial motion from a 3D navigational perspective based on the virtual interface. In addition, geometric collision and dangerous proximity of tool-tissue distance can be computed for automatic safety management.

Earlier works [63, 65] involving an active approach requires an additional incision port or fiducial placement. More recent works [50, 66] in the robotics aspect use a passive approach where natural features are exploited for tracking. This is a preferred clinical setting due to its streamlined framework without having to introduce an additional tracking sensor or the need to perform inter-sensor registration.

A futuristic but promising aspect of the surgical robotic navigation system is its role in Natural Orifice Transluminal Endoscopic Surgery (NOTES). Low et al. [71] reported the feasibility of NOTES application with a master-slave surgical robotic system (Fig. 13) originally used for the therapeutic gastrointestinal endoscopic procedure. The robot-assisted surgical system was later used to perform Robot-Assisted Endoscopic Submucosal Dissection [72]. The study was demonstrated in vitro experiments. Wireless controlled miniature surgical robots can also be endoscopically placed into the peritoneal cavity. Such technology uses onboard cameras for visual feedback and navigation.

4.3 Augmented reality visual guidance

The increasingly complex minimally invasive procedure has spawned the development of surgical navigational guidance towards a more user-centric system. Traditional image-guided surgery has, over the years, evolved to include intuitive interface equipped with powerful augmented reality (AR) platform. A vision-based technique is a contributive factor to this development. In fact, an AR overlay of tumour onto the liver was demonstrated using MC-SLAM discussed previously [60] that take into account the deformation model of the liver surface due to respiration. While a mean error of 0.11 cm in ex vivo study was reported for deformation tracking, no accuracy evaluation of the AR visualisation in vivo was presented for that study. Case Study VI features one of the complete AR systems in terms of functionality.
Case Study VI [RA23]: AR during robot-assisted laparoscopic partial nephrectomy

**Background and rationale:** A minimally invasive surgical treatment for renal cell carcinoma is through laparoscopic partial nephrectomy (LPN). The localisation and identification of tumour boundary is an important but difficult aspect of the procedure. It is essential to remove tumour totally while maximising on the nephrons preserve based on pre-operative information and laparoscopic guidance. While the use of ultrasound supports intra-operative guidance, the visualisation and acquisition ability is limited.

**Method:** In the work proposed by Su et al. [70], vision-based tracking is used for continuous registration of the virtual model and overlaying of the pre-operative diagnostic model onto the actual surgical scene. The process, as shown in Fig. 14 includes manual segmentation of CT image to generate 3D surface anatomical models, calibration of recorded stereoscopic video to 3D CT model, surface-based tracking through triangulation of 3D points, and registration refinement using 3D–3D modified iterative closest point (ICP) registration.

**Contribution:** In essence, this study implemented augmented reality on stereo-endoscopic images for real-time visualisation of pre-operative CT imaging, image-model registration and overlay with accuracy within 1 mm. Despite it being semi-automated, the study demonstrated excellent potential for effective surgical navigation and represented the state-of-the-art technology in AR for vision-based tracking surgical navigation in terms of functionalities.

**Comment:** While the system demonstrated full-fledge AR functionality, evaluation of the reliability and efficacy of the system was not reported. There has not been a substantial study to validate its clinical value. For the application to further extend to other laparoscopic surgeries, for example liver surgery, deformable registration will have to be considered. It was also mentioned in the article under its comment section that the system can only track gross motion but not deformation. The authors also acknowledge that the current limitation of the system lies in its inability to manage anatomical deformation, which they agree is a significant factor affecting the accuracy. They further highlighted the problem of tumour displacement in liver surgery as a significant factor that decreases the precision of most surgical navigation.

The work in RA7 by Grasa et al. [51] is a more balanced coverage of both the AR functionality as well as the theoretical treatment of endoscope and anatomical features tracking for surgical navigation. The AR features of this work include scene annotation with dimension information and point-to-point distance measurement. In RA24, an integrated visualisation platform for navigational view during fetoscopic procedures is introduced [56]. The system is designed to be self-contained using only the very information from the fetoscopic views. The workflow is shown in Fig. 15.

The results of the vision-based self-contained approach is a dynamic view expansion of the limited fetoscopic view with an augmented cue of the position of the fetoscopic instrument with respect to the surrounding, hence providing intuitive navigational guidance as illustrated in Fig. 16.

5 Discussion

5.1 Significance and impact

Interests in vision-based tracking in surgical navigation are driven by many factors. Limitations and challenges faced by other forms of navigation systems are a strong motivation for exploring alternative approaches through vision-based tracking. Typical surgical navigation systems perform tracking application of surgical tools with external tracking systems. This usually involves the attachment of a bulky sensor to the surgical tools and complicated cross-domain registration of information from the non-image-based sensor to the visual guidance in surgical navigation. Therefore, vision-based tracking is believed to offer a
more streamlined framework compared to odometry-based sensing, including gyro and accelerometer. Other external trackers like a magnetic sensor or optical tracker face problems in magnetically induced noise and occlusion, respectively.

Development in vision-based surgical navigation has a great impact on accelerating the trend of minimally invasive techniques. Prominent examples include research efforts in NOTES technology development [73]. NOTES is a futuristic approach where no incision outside the patient's body is required. It can be done through the insertion of a flexible robotic endoscope creating an incision on the gastric wall to access the abdominal cavity leaving no scar on the patient's body. Like other flexible endoscopes, the external tracker is not a viable option in most cases due to the non-rigid multi-segmented endoscope. A vision-based approach appears to be the most practical tracking technique for navigation in NOTES.

### 5.2 Challenges

Technical challenges in vision-based tracking for surgical navigation are largely due to the complexity of the field of view associated with inter-lumen reflection, specular highlight and the general lack of stable distinctive features [74]. This leads to ill-conditioned problems and the inability to impose well defined geometrical constraints. In addition, surgical scenes are often subjected to dynamic conditions like free-form deformation making vision-based tracking challenging.

While the vision-based technique for surgical navigation has demonstrated promising potential and contributions to the research scene of computer-aided surgery, its practical implementation for actual clinical practice remains limited. A major challenge in clinical translation is the issue of validation and fulfilling regulatory requirements. While the topic has been widely researched upon, a standard for system precision, accuracy and reliability is not well established. There exists a knowledge-gap in establishing a scientifically sound methodology for system validation and measurement of system reliability. This knowledge-gap is a serious bottleneck for vision-based surgical navigation to exhibit its fullest potential. The nature of the approach has made validation challenging due to a lack of ground truth for in vivo studies. Although promising data have been reported in the literature suggesting the feasibility and efficacy of a vision-based approach, it is not supported by strong scientific quantification of most of the time. For instance, a study on the RE-VE registration approach [63] can only report accuracy based on visual inspection. Its reliability was also based on an anecdotic claim that ‘its best case produces 600 consecutive successfully tracked frames’. Similarly, in quantifying the error for registration, it is difficult to isolate subjective elements. The concept of true registration error (TRE) is difficult to use for comparison across studies as it is complicated by inter-subject factors. For example, the TRE was determined by the task execution of surgeons reported in [75].

Despite presenting an attractive streamlined implementation and operation frameworks, pure vision-based tracking is ultimately limited by the capacity of the information source. Solely vision-acquired data has limitations. A typical issue, for example is that vision-based pose estimation is usually computed sequentially with incremental frames that lead to drift in estimation from an absolute frame reference. Moreover, projective geometry only estimates the spatial position up to a scale factor. This calls for the incorporation of information sources from other sensing systems. Sophisticated medical imaging modalities like CT scan and magnetic resonance imaging are technically difficult to implement in real-time during surgery. Works involving the fusion of pre-operation imaging data with endoscopic images are proposed [47, 52, 53, 57, 70] but faces limitations in constructing timely anatomical map as it relies only on intra-operative endoscopic updates of the surgical scene. While there is work that proposed Bayesian fusion framework, these studies [33, 41] are nevertheless based solely on the single information source from the endoscope. Significant works that fuse multi-sensor data for vision-based tracking in surgical navigation are limited.

### 5.3 Future trends

Various vision-based tracking methods for surgical navigation have been proposed and it seems that the next phase of development in the field is the actual implementation of clinical practice. This process is known as clinical translation in the context of R&D in the medical industry. It appears that to ensure clinical translation of the vision-based surgical navigation research effort into practical use, the scientific methodology in validation must be further looked into so as to overcome the regulatory barrier discussed earlier on.

A highly possible trend in the development of vision-based tracking on surgical navigation is to tap on the advancement of other technologies like that in electronics, optics, and imaging. Biophotonic imaging approaches have received much attention recently [6]. In addition, with improvement in the optical and electronic devices, depth cameras in the form of stereoscopic or infrared depth-sensing may be more ubiquitous and feasible for incorporating into existing endoscopic instruments. The development of sophisticated, active approaches like projected coded patterns and time-of-flight technologies [5] for robustness against the dynamic environment is likely to be leveraged. Multimodality imaging and information fusion have been a successful approach in the domain of the medical image. As mentioned in this paper, studies on such an approach are limited in current literature. A good example has, however, been demonstrated by Mori et al. [47]. It has been shown that the timing performance of using a magnetic sensor is much better than that of his previous study [52]. Future developments are likely to involve the fusion of information from a multisensory system.

Artificial intelligence could be a promising trend that overcomes some of the mentioned challenges. With the consistently increasing computational capability, surgical navigation may move into a new paradigm using an extensive data-driven approach. However, consideration might not be merely technical. Regulatory issues and concerns about policy-making may play a significant role in shaping the next-generation surgical navigation as well.
6 Conclusion

In this review, contemporary research and development were reviewed. The technology level of this subject was also discussed. Based on the material reviewed and discussed, original views on the technology’s significance, challenges and future developmental trend were presented. Through this review exercise, the author gained invaluable insights relevant to his research topic on medical image mapping and information fusion for surgical navigation. The research is set to focus on clinical applications specific to foetal surgery when pre-operative imaging data is usually limited to ultrasound scanning. The aim is to provide a surgeon with a navigational perspective through registration of the 2D endoscopic image to the 3D ultrasound model.

Limitation in external tracking has inspired the author to adopt a vision-based technique complemented with data fusion from ultrasound imaging. The incorporation of positional data of a vision-based technique complemented with data fusion from ultrasound enhance the image mapping and information fusion for surgical navigation. The technology level of this subject was also discussed. This is an open access article published by the IET and Zhejiang University Press under the Creative Commons Attribution License http://creativecommons.org/licenses/by/3.0/)

7 Reviewed articles

See Table 3.

| S/n | Lead author | Year | Title |
|-----|-------------|------|-------|
| RA1 | Mori | 2002 | Tracking of a bronchoscope using epipolar geometry analysis and intensity-based image registration of real and virtual endoscopic images [47] |
| RA2 | Mori | 2005 | Hybrid Bronchoscope Tracking Using a Magnetic Tracking Sensor and Image Registration [52] |
| RA3 | Deguchi | 2009 | Selective registration measure for bronchoscope tracking based on image registration [53] |
| RA4 | Burschka | 2005 | Scale-invariant representation of monocular endoscopic images to CT-scans for sinus surgery [57] |
| RA5 | Mountney | 2006 | Simultaneous Stereoscope Localisation and Soft-Tissue Mapping for Minimal Invasive Surgery [46] |
| RA6 | Noonan | 2009 | Stereoscopic fibroscope for camera motion and 3D depth recovery during Minimally Invasive Surgery [58] |
| RA7 | Grasa | 2009 | EKF Monocular SLAM 3D modelling, Measuring and Augmented Reality from Endoscope Image Sequences [51] |
| RA8 | Grasa | 2011 | EKF monocular SLAM with relocalisation for laparoscopic sequences [48] |
| RA9 | Mountney | 2010 | Motion Compensated SLAM for Image-Guided Surgery [60] |
| RA10 | Nakamura | 2001 | Heartbeat synchronisation for robotic cardiac surgery [63] |
| RA11 | Hayashibe | 2006 | Laser-scan endoscope system for intraoperative geometry acquisition and surgical robot safety management [65] |
| RA12 | Mountney | 2007 | Probabilistic Framework for Tracking Deformable Soft Tissue in Minimally Invasive Surgery [41] |
| RA13 | Lo | 2009 | Belief Propagation for Depth Cue Fusion in Minimally Invasive Surgery [33] |
| RA14 | Stoyanov | 2009 | Soft tissue deformation tracking for robotic-assisted minimally invasive surgery [36] |
| RA15 | Mountney | 2008 | Soft Tissue Tracking for Minimally Invasive Surgery: Learning Local Deformation Online [61] |
| RA16 | Richa | 2010 | Beating heart motion prediction for robust visual tracking [49] |
| RA17 | Richa | 2008 | Efficient 3D Tracking for Motion Compensation in Beating Heart Surgery [34] |
| RA18 | Lau | 2004 | Stereo-Based Endoscopic Tracking of Cardiac Surface Deformation [76] |
| RA19 | Richa | 2008 | Deformable motion tracking of the heart surface [68] |
| RA20 | Richa | 2010 | Three-dimensional Motion Tracking for Beating Heart Surgery Using a Thin-plate Spline Deformable Model [37] |
| RA21 | Richa | 2010 | Robust 3D Visual Tracking for Robotic-Assisted Cardiac Interventions [50] |
| RA22 | Yang | 2010 | A robotic system for overlapping radiofrequency ablation in large tumour treatment [54] |
| RA23 | Yang | 2015 | Vision-based endoscope tracking for 3D ultrasound image-guided surgical navigation [55] |
| RA24 | Yang | 2015 | Self-contained image mapping of placental vasculature in 3D ultrasound-guided fetoscopy [56] |
| RA25 | Fischer | 2006 | Hybrid tracking method for surgical augmented reality [77] |
| RA26 | Su | 2009 | Augmented Reality During Robot-assisted Laparoscopic Partial Nephrectomy [78] |
| RA27 | Wen | 2013 | Hand gesture guided robot-assisted surgery based on a direct augmented reality interface [79] |
| RA28 | Wang | 2014 | Augmented Reality Navigation With Automatic Marker-Free Image Registration Using 3D Image Overlay for Dental Surgery [80] |
| RA29 | Wang | 2016 | Video see-through augmented reality for oral and maxillofacial surgery [81] |
| RA30 | Shvets | 2018 | Automatic Instrument Segmentation in Robot-Assisted Surgery using Deep Learning [67] |

8 References

[1] Sylvester, D., Sullivan, S.M.: ‘Virtual surgical planning (VSP)’, in ‘Craniofacial 3D imaging’ (Springer, Cham, Switzerland, 2019), pp. 227–236

[2] Trusty, P.M., Slesnick, T.C., Wei, Z.A., et al.: ‘Fontan surgical planning: previous accomplishments, current challenges, and future directions’, J. Cardiovasc. Transl. Res., 2018, 11, (2), pp. 133–144

[3] Yuan, P., Mai, H., Li, l., et al.: ‘Design, development and clinical validation of computer-aided surgical simulation system for streamlined orthognathic surgical planning’, Int. J. Comput. Assist. Radiol. Surg., 2017, 12, (12), pp. 2129–2143

[4] Liu, P., Qin, J., Duan, B., et al.: ‘Overlapping radiofrequency ablation planning and robot – assisted needle insertion for large liver tumors’, Int. J. Med. Robot., 2019, 15, (1), p. e1692

[5] Mirota, D.J., Iishi, M., Hager, G.D.: ‘Vision-based navigation in image-guided interventions’, Ann. Rev. Biomed. Eng., 2011, 13, (1), pp. 289–297

[6] Stoyanov, D.: ‘Surgical vision’, Ann. Biomed. Eng., 2012, 40, (2), pp. 332–345

[7] Modi, Y.S., Ehlers, J.P.: ‘Heads-up vitreoretinal surgery: emerging technology in surgical visualization’, Retinal Physician, 2016, 13, pp. 26–29

[8] Fan, Z., Weng, Y., Chen, G., et al.: ‘3D interactive surgical visualization system using mobile spatial information acquisition and autostereoscopic display’, J. Biomed. Inf., 2017, 71, pp. 154–164

[9] Davies, B.: ‘Robots in minimally invasive surgery’. IEEE Colloquium on Through the Keyhole: Microengineering in Minimally Invasive Surgery, London, UK., 1995, pp. 5/1–5/2

[10] Lee, S.-L., Lerotic, M., Vitello, V., et al.: ‘From medical images to minimally invasive intervention: computer assistance for robotic surgery’, Comput. Med. Imaging Graph., 2010, 34, (1), pp. 33–45

[11] Vitello, V., Lee, S.-L., Candy, T.P., et al.: ‘Emerging robotic platforms for minimally invasive surgery’, IEEE Rev. Biomed. Eng., 2013, 6, pp. 111–126

[12] Abdelal, A.E., Mathur, P., Salcedan, S.E.: ‘Robotics in vivo: a perspective on human–robot interaction in surgical robotics’, Ann. Rev. Control Robot. Auton. Syst., 2020, 3, pp. 221–242

[13] Panesar, S., Cagle, Y., Chander, D., et al.: ‘Artificial intelligence and the future of surgical robotics’, Ann. Surg., 2019, 270, (2), pp. 223–226

[14] Chang, S.K.Y., Hlaing, W.W., Yang, L., et al.: ‘Current technology in navigation and robotics for liver tumours ablation’, Ann. Acad. Med. Singapore, 2011, 40, (5), pp. 231–236
[74] Delingette, H., Pennec, X., Soler, L., et al.: ‘Computational models for image-guided robot-assisted and simulated medical interventions’, Proc. IEEE, 2006, 94, (9), pp. 1678–1688

[75] Helferty, J.P., Higgins, W.E.: ‘Technique for registering 3D virtual CT images to endoscopic video’. 2001 Proc. Int. Conf. on Image Processing, 2001, Thessaloniki, Greece, 2001, vol. 2, pp. 893–896

[76] Lau, W.W., Ramey, N.A., Corso, J.J., et al.: ‘Stereo-based endoscopic tracking of cardiac surface deformation’. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Saint-Malo, France, 2004, pp. 494–501

[77] Fischer, J., Eichler, M., Bartz, D., et al.: ‘A hybrid tracking method for surgical augmented reality’, Comput. Graph., 2007, 31, (1), pp. 39–52

[78] Su, L.-M., Vagvolgyi, B.P., Agarwal, R., et al.: ‘Augmented reality during robot-assisted laparoscopic partial nephrectomy: toward real-time 3D-CT to stereoscopic video registration’. Urology, 2009, 73, (4), pp. 896–900

[79] Wen, R., Tay, W.L., Nguyen, B.P. et al.: ‘Hand gesture guided robot-assisted surgery based on a direct augmented reality interface’, Comput. Methods Programs Biomed., 2014, 116, (2), pp. 68–80

[80] Wang, J., Suenaga, H., Hoshi, K., et al.: ‘Augmented reality navigation with automatic marker-free image registration using 3-D image overlay for dental surgery’, IEEE Trans. Biomed. Eng., 2014, 61, (4), pp. 1295–1304

[81] Wang, J., Suenaga, H., Yang, L., et al.: ‘Video see-through augmented reality for oral and maxillofacial surgery’, Int. J. Med. Robot. Comput. Assist. Surg., 2017, 13, (2), p. e1754