Gaze gesture based human robot interaction for laparoscopic surgery✩

Kenko Fujii1, Gauthier Gras1, Antonino Salerno, Guang-Zhong Yang∗

The Hamlyn Centre for Robotic Surgery, Imperial College London, UK

A R T I C L E   I N F O

Article history:
Received 12 December 2016
Revised 22 November 2017
Accepted 23 November 2017
Available online 28 November 2017

A B S T R A C T

While minimally invasive surgery offers great benefits in terms of reduced patient trauma, bleeding, as well as faster recovery time, it still presents surgeons with major ergonomic challenges. Laparoscopic surgery requires the surgeon to bimanually control surgical instruments during the operation. A dedicated assistant is thus required to manoeuvre the camera, which is often difficult to synchronise with the surgeon's movements. This article introduces a robotic system in which a rigid endoscope held by a robotic arm is controlled via the surgeon's eye movement, thus forgoing the need for a camera assistant. Gaze gestures detected via a series of eye movements are used to convey the surgeon's intention to initiate gaze contingent camera control. Hidden Markov Models (HMMs) are used for real-time gaze gesture recognition, allowing the robotic camera to pan, tilt, and zoom, whilst immune to aberrant or unintentional eye movements. A novel online calibration method for the gaze tracker is proposed, which overcomes calibration drift and simplifies its clinical application. This robotic system has been validated by comprehensive user trials and a detailed analysis performed on usability metrics to assess the performance of the system. The results demonstrate that the surgeons can perform their tasks quicker and more efficiently when compared to the use of a camera assistant or foot switches.

© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license. (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Technological advances over the past decade have enabled the routine use of Minimally Invasive Surgery (MIS) in an increasing number of clinical specialities. MIS offers several benefits to patients including a reduction in operating trauma, post-operative pain, and faster recovery times. It has also led to budgetary benefits for hospitals through cost savings from reduced hospitalisation duration.

Performing laparoscopic surgery requires bimanual manipulation of surgical instruments by the surgeon. The field-of-view (FOV) of laparoscopic cameras is usually very narrow. In order to assist with the navigation during the operation, a surgical assistant usually manoeuvres the laparoscope camera on behalf of the operating surgeon. Understanding the surgeon's desired FOV, and communication via verbal instruction can be challenging. Failure to provide good visualisation of the operating field not only induces greater mental workload on the surgeon (Zheng et al., 2012b), but also leads to unrecognised collateral injuries (Nduka et al., 1994). The need for good camera handling has been recognised as an important step in the training curricula for surgical residents (Shetty et al., 2012). In spite of such training, issues related to assistant fatigue, hand tremor and the confined shared workspace for the surgeon and camera assistant persist. There are also man-power and cost implications to requiring a highly skilled camera assistant, which if addressed, could allow a surgeon to operate solo, thereby improving cost and staffing efficiency. In order to address the above deficiencies, a number of commercial robotic assisted camera systems have been developed. These include the EndoAssist (Kommu et al., 2007), which is controlled by the user's head-mounted infrared emitter, the verbally controlled Automatic Endoscope Optimal Position (AESOP) system (Kraft et al., 2004), and the finger stick controlled SoloAssist from AktorMed (Sriskantharajah et al., 2012).

Gaze information obtained by a remote eye tracker, i.e. where the person is looking, can be used to create a gaze contingent control system to move the camera. With this gaze contingent control, it is possible to move the camera by following the user's gaze position on the target anatomy. A number of gaze contingent robotic assisted camera systems (Noonan et al., 2010; Staub et al., 2012) have been developed, as well as methods using gaze data to recover 3D fixation points and perform camera motion stabilisation (Mylonas et al., 2006). Existing gaze contingent camera systems, however, only have a panning control of the laparoscope. The lack of zoom and tilt control complicates effective navigation during surgery. The system proposed by Noonan, for example, requires a
foot-pedal which the user needs to press to activate. The need to introduce additional hardware such as foot-pedals can lead to instrument clutter in an already complex environment. Furthermore, existing eye-controlled platforms often use dwell time on fixed regions to indicate a user’s intention, which can be difficult to use in practice. Several methods have been developed using gaze data as a central component for intention recognition in a robotic system, for automatic laser targeting (Gras and Yang, 2016) and adaptive motion scaling (Gras et al., 2017).

In order to overcome these problems, this article introduces a gaze contingent robotic control system where the camera is activated via real-time gaze gestures rather than an external switch. Through the use of multiple gestures, which are statistically learned and can map to specific camera control commands, we show that it is possible to use different camera control modes such as panning, zooming, and tilting without interfering with the user’s natural visual search behaviour. The proposed system also incorporates a novel online calibration algorithm for the gaze tracker overcoming the need of an explicit offline calibration procedure. The proposed gaze gesture based human-computer interaction method differs from previous gaze based interaction methods such as the eye mouse (Jacob, 1991), which uses dwell time to convey user intention of mouse clicking, or the Manual And Gaze Input Cascaded (MAGIC) pointing method (Zhai et al., 1999), which moves the cursor position in close proximity to the target location, but relies on the user to convey their intention with a manual cursor movement and mouse click. The work presented here builds on the “Perceptual Docking” paradigm introduced by Yang et al. (2008), and extends initial work presented in Fujii et al. (2013). The key novelties of the work presented include: i) the capability to pan, zoom, and tilt the camera (rotation around the laparoscope’s longitudinal axis); ii) the ability to seamlessly switch between panning and zooming control by using the distance from the user’s face to convey the intention to zoom in or out; and iii) an implicit online calibration method for the gaze tracker, which overcomes the need of an explicit offline gaze calibration before using the gaze contingent system, offering a fast, frustration-free user experience of the system. Furthermore, an exhaustive evaluation of these novelties is presented, drawing data from extensive user studies.

The article is organised as follows; Section 2.1 presents the gaze contingent laparoscope system, including the system design and implementation. The proposed online calibration method is detailed in Section 2.2. Finally, Sections 3.1 and 3.2 present a detailed evaluation of the gaze contingent laparoscopic control and the online calibration respectively.

2. Methods

2.1. Gaze contingent laparoscopic control

2.1.1. System overview

The aim of the proposed system is to provide an interface that will enable hands-free camera activation, allowing the surgeon to perform a bimanual task without the need for a camera assistant. Furthermore, in order to limit the cognitive burden on the surgeon, this interface must function without requiring additional foot-pedal hardware or using gaze dwell-time methods. This is achieved through the use of gaze gestures. Gaze gestures are based on fast eye movements, i.e. saccadic movements rather than fixations. They consist of a predefined sequence of eye movements. Gaze gestures can be single-stroke (Drewes et al., 2007; Mollenbach et al., 2008) or multi-stroke (Rozado et al., 2012). When the user performs the intended unique sequence of saccadic eye movements, a specific command is activated. The use of gaze gestures have previously been applied to eye typing (Wobbrock et al., 2008), Human-Computer Interaction, mobile phone interaction (Rozado et al., 2013) and gaming (Istance et al., 2010). The key components of the proposed system are illustrated in Fig. 1. It comprises of a Tobii 1750 remote gaze tracker, a Kuka Light Weight Robot (LWR) (KUKA Roboter GmbH, Augsburg, Germany), a 10 mm zero degree Karl Storz rigid endoscope, and a Storz Tele Pack light box and camera. Two Storz Matkowitz grasping forceps and an upper gastrointestinal phantom with simulated white lesions were used for the evaluation of the system. Additionally, the laparoscope and surgical tools were tracked using an NDI Polaris Vicra infrared tracker during the experiments (NDI Medical, Ontario, Canada).

The human eye is normally used for information gathering (Yang et al., 2002) rather than to convey intention to control external devices. As such, the main challenge of using gaze gestures is distinguishing natural gaze patterns from intentional gaze gestures with high accuracy and precision. To this end, pattern recognition methods are necessary to learn these gaze gestures. The proposed system uses gaze gesture recognition based on Hidden Markov Model (HMM) to learn multiple input commands from a surgeon in order to convey the desired camera control mode.

Two possible gaze gestures are introduced to control the camera: activate camera (for pan and zoom) and tilt camera (for rotation). The activate camera gaze gesture is illustrated in Fig. 2(a). It is defined by the following three-stroke sequence of eye movements: gaze at the centre of the screen, then to the bottom right corner, then back to the centre, and finally back to the bottom right corner. The tilt camera gaze gesture is similar to the activate camera gaze gesture but the user is instead required to look at the bottom left corner of the screen, as shown in Fig. 2(b). Gaze gestures oriented towards the corner of the screen are chosen to prevent obstruction of the camera view, minimise the amount of screen
space necessary, and reduce detection of involuntary gaze gestures. Text labels were placed at the bottom left and right corners of the screen to indicate the mode to be activated by performing a gaze gesture in that direction.

Once a gaze gesture is identified, the robotic arm is activated and the user is able to control the laparoscope with his gaze. The user can also deactivate the gaze contingent laparoscopic control via gaze gesture. The control mechanism of the system is represented in Fig. 3. It is composed of two processes: a gaze gesture recognition process and a robot control process. The gaze gesture recognition process analyses inbound gaze data and identifies whether a gesture has been performed. The robot control process uses the Point-of-Regard (PoR) data, i.e., where the user is looking at, to generate the robot trajectory. Finally, the user can stop the robotic camera control by fixing the stop camera text present at the bottom left corner of the screen during robotic control.

2.1.2. Gesture recognition

The PoR gaze data generated by the eye tracker is first passed through a median filter with a 150ms time window to reduce the noise inherent to eye data. The filtered data is then used in a gesture segmentation algorithm. Since the proposed gaze gestures consist of three fast eye movements, i.e., saccades, the gesture segmentation algorithm was designed to detect these sequential saccades. These saccades are detected by searching for two sequential gaze data samples which exceed a 300° s⁻¹ velocity threshold. In this context, the term “stroke” is also used to describe a saccade. A velocity threshold saccade detection technique (Salvucci and Goldberg, 2000) was implemented. Additional temporal constraints were added to ensure that the gaze gestures were not excessively long (each gaze stroke was required to be less than 750 ms), allowing segmentation of continuous and user-intended gaze gestures. This segmentation process outputs a stream of strokes which is analysed by the gesture recognizer in a sliding window fashion. Every new stroke detected is analysed alongside the last two strokes stored in memory to identify a potential gesture. In case of failure, the oldest stroke is discarded, and the two others are stored until a new stroke is detected and the process can be repeated. The buffer is cleared and the search reset when a gaze gesture is recognised. The xy coordinates from the segmented trajectories of a potential gaze gesture are then clustered using a pre-trained k-means algorithm. Each cluster’s symbol number, centroid coordinates, and radius are used collectively to create a discrete codebook that captures the relevant features of the gaze gestures, i.e., the xy coordinates. The codebook was designed offline from 600 gaze gesture training data sequences, and five clusters were chosen for the k-means algorithm. Each potential gaze gesture sequence is encoded using the codebook, where symbol numbers are assigned to each observation (xy coordinate) by using the distance between the observation and the centroid of each cluster, provided that it is within the defined radius. If an observation is outside the feature space, it is discarded.

In order to recognise the segmented potential gaze gestures, two left-to-right HMMs were used for each camera control activation mode. Unlike in Mollenbach et al. (2009), gaze patterns are not analysed here just to identify the nature of the gaze data, i.e., whether it constitutes a continuous motion, fixation or a stroke, but to classify which kind of gesture is being performed. The activate camera gaze gesture is modelled by HMM1 and enables panning and zooming control of the laparoscope. The tilt camera gaze gesture is modelled by HMM2 and enables rotation around the laparoscope’s longitudinal axis.

2.1.3. Model training

HMMs are stochastic sequences where the hidden states are associated with a probability (Rabiner, 1989) and are typically represented by a set of N interconnected states S = [S₁, …, S₇], as well as a number K of discrete symbols V = [v₁, …, vₖ]. The model
is described by a sequence of $M$ observations, a state transition matrix $E = [e_{ij}]$ and an emission probability matrix is $F = [f_{jk}]$.

The state transition matrix $E$ describes transition probabilities from state $i$ to state $j$ as $e_{ij} = P(q_{t+1} = S_j | q_t = S_i)$, where $1 \leq i, j \leq N$, and $q_t$ is the state at time $t$. The emission probability matrix $F$ describes the probabilities $f_{jk}$ of generating the symbol $v_t$ from the state $q_t$ with $f_{jk} = P(v_t = S_j | q_t = S_i)$. $1 \leq j \leq N$ and $1 \leq k \leq K$. Additionally, initial state probabilities must be selected to initialise the model.

Each of the HMMs model parameters was trained offline using a set of gaze gesture training data. During training, both intentional and unintentional gaze gesture data sequences were included in our data sets. The data sets were collected from twenty participants who did not participate in usability trials of the gaze contingent laparoscope system. Participants performed a gaze calibration procedure prior to the data collection, and the accuracy of the calibration was verified. Each participant provided 30 repetitions of each of the two types of intentional gaze gestures. The task during the gaze gesture data collection was to perform the three-legged gaze gestures whilst observing a black screen with white guidance dots in the middle and in the lower two corners of the screen. Participants were asked to perform a gaze gesture starting from the middle, moving to the corner, back to the centre and the back to the corner. The resulting training data consisted of 600 intentional gaze gestures for each HMM. Additionally, unintentional gaze gesture data was collected during a five-minute web browsing task. More specifically, the unintended gaze gesture data were collected whilst viewing a number of websites which consisted of image based content. Subjects were asked to spend five-minutes browsing the site while eye tracking data was being recorded in the background. The intention was to simulate random gaze behaviour. The data collected during this task was used in the learning phase of the HMMs in order to improve robustness of the HMMs against false positives.

Each of the 600 intentional gaze gesture training sequences was encoded using the formulated k-means clustering codebook. An initial state probability is defined within this set of training data observations, and optimal state transition and emission probabilities that describe the set of training observations are iteratively obtained using the Baum–Welch algorithm. The initial state probabilities were randomly initialised between 0 and 1. To improve the trade-off between sensitivity, false positive rate and overall complexity of the system, a 10-fold cross validation was run across the HMM with different number of states (between 2 and 8 states, total of seven runs as shown in Fig. 4(a–b)). 90% of the encoded training sequences are used with the Baum–Welch algorithm to iteratively obtain the HMMs parameters. With each of the seven training sets, a detection probability threshold was set at a 95% confidence limit of the training data sequences’ inference values, i.e. the probabilities of these sequences given the trained HMM. The recognition accuracy of the HMM is then defined by using this threshold on the remaining 10%, the validation data set. A six state HMM with an inference threshold of 0.7 was found to provide the best overall performance for both the activate camera and tilt camera gaze gesture detection. The rational of choosing this threshold value is apparent in the receiver operating characteristic (ROC) curve of the 6 state HMM which is illustrated in Fig. 4(c), with the respective close up version shown in Fig. 4(d). From these figures, it is observable that the threshold of 0.7 provides the best trade off of sensitivity and false positive rate; with a sensitivity of 0.98 and false positive rate of 0.01 for the activate camera (HMM1) and 0.98 and 0.02 respectively for the tilt camera (HMM2). The ROC curve illustrates the robustness of the three-stroke gaze gestures toward the edge of the screen; there is very little overlap between the visual search behavioural noise and the gaze gestures. Inference value histograms obtained from testing the 6 state HMMs with the

![Fig. 4. HMM training 10-fold cross validation for the (a) activate camera and (b) tilt camera HMM. The ROC curve (c) for both gaze gestures (d) both gaze gestures zoomed view. The inference values of intended gaze gestures for (e) activate camera gestures and (f) tilt camera gestures. The inference values of activate camera and tilt camera gestures given the use of the activate camera gesture trained HMM (g) and the inference values of tilt camera and activate camera gestures given the use of the tilt camera gesture trained HMM (h).](image-url)
to update a minimum jerk trajectory planner providing the desired pose $x_d$ to the robot controller based on a Cartesian impedance control scheme. The controller computes the command torque $\tau_d$ for each joint according to the compensation of the whole robot dynamics and the error signal $e$ estimated as the difference between the reference pose $x_d$ and the actual pose $x$ in Cartesian space retrieved by the robot's forward kinematics. The Cartesian impedance control has been chosen to guarantee both a safe human robot interaction and an intuitive camera positioning during surgery. Details about the impedance control law can be found in Albu-Schaffer et al. (2007a; 2007b).

The coordinates $P_{\text{C}} = [x_{\text{eye}} y_{\text{eye}} z_{\text{eye}}]^T$ are expressed in the $xy$ plane of the camera frame (C), corresponding to the eye-trackers screen. The angle $\alpha$ of the next camera motion in the same plane can be reconstructed as

$$\alpha = \tan^{-1}\left(\frac{y_{\text{eye}}}{x_{\text{eye}}}\right). \quad (1)$$

The proposed system computes the final position $p_f^c$ with respect to the camera frame (C) by considering a simultaneous pan and zoom motion as

$$p_f^c = \begin{bmatrix} L_1 \cos(\alpha) & L_1 \sin(\alpha) \end{bmatrix}^T, \quad (2)$$

where $L_1$ and $L_2$ are the displacement in the plane $xy$ (i.e., pan motion) and in the $z$-axis (i.e., zoom motion) of the camera frame (C) respectively. If only the pan mode is active then $L_2 = 0$. In the case of a strictly zooming out motion, $\alpha = 0$ and the equation simplifies to

$$p_f^c = \begin{bmatrix} 0 & 0 & L_1 \end{bmatrix}^T. \quad (3)$$

The final position $p_f^b$ expressed in the base frame of the robot (B) is given by:

$$p_f^b = \begin{bmatrix} 1 \end{bmatrix}^T \begin{bmatrix} p_f^c \end{bmatrix}^T = T_b^c \begin{bmatrix} 1 \end{bmatrix}^T, \quad (4)$$

where $T_b^c$ is the homogeneous transformation matrix of the camera frame (C) with respect to the base frame of the robot (B). Robot motions are generated from these control points using a minimum-jerk (degree 5 polynomial) interpolation to provide a smooth trajectory. A remote control of motion (RCM) constraint is also used to force the robot to respect the trocar through which the laparoscope is inserted. Let $C$ be the position of the RCM constraint relative to the base robot frame (B). Given the vector $v = p_f^b - p_c^b$ and the $z$-axis unit vector $z_c$ of the camera frame (C), a rotation matrix $R(\beta, \hat{\alpha})$ can be defined to constrain the orientation of the laparoscope, so that it passes through $G$ during the motion from an initial point $p_f^b$ to $p_f^c$. $R(\beta, \hat{\alpha})$ is defined as the matrix corresponding to the angle-axis rotation defined by the rotation angle $\beta$, of final value $\beta_f$, and the rotation axis $\hat{\alpha}$, where:

$$\beta_f = \tan^{-1}\left(\frac{\|z_c \times v\|}{\|z_c\| \|v\|}\right), \quad \hat{\alpha} = \frac{z_c \times v}{\|z_c \times v\|}, \quad (5)$$

Accordingly:

$$\begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix} \begin{bmatrix} p_f^b \end{bmatrix} = \begin{bmatrix} p_f^c \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix} \end{bmatrix} \begin{bmatrix} p_f^b \end{bmatrix} = \begin{bmatrix} p_f^c \end{bmatrix} \quad (6)$$

where $\begin{bmatrix} \begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix} \end{bmatrix}$ is the homogeneous transformation of the camera frame (C) expressed in the base robot frame (B) at the beginning of a motion. $\begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix}$ is that same matrix after having completed the motion and enforced the RCM, $q$ represents the robot joint values, and $\begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix}$ is the initial rotation matrix of the camera frame (C) relative to the base robot frame (B). Like the control points, $\beta$ follows a minimum-jerk interpolation between 0 and $\beta_f$ to simultaneously rotate and translate the camera frame.

In order to perform the tilt camera motion, a rotation $\gamma$ around the longitudinal axis of the laparoscope $z_c$ is added to $\begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix}$ in Eq. (6). Additionally, $L_4 = 0$ and $L_5 = 0$ as there is no panning or zooming motion in this mode. $\begin{bmatrix} \begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix} \end{bmatrix}$ can be re-written as:

$$\begin{bmatrix} \begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} R(\beta, \hat{\alpha}) \end{bmatrix} \end{bmatrix} \begin{bmatrix} 1 \end{bmatrix} \quad (7)$$

where $R(\gamma)$ represents the elementary rotation matrix of the frame (C) about the unitary vector $z_c$.

2.1.5. Interface design

The implemented control User Interface (UI) is illustrated in Fig. 5. On system initialisation, the camera is stationary and the system waits for a gaze gesture input from the user. The user has the option to control the camera via activate camera or tilt camera modes. Activate camera mode enables panning or zooming. It is activated by one gaze gesture and switching between the panning and zooming is enabled with a movement of the head forward or backward. This provides a combined pan and zoom control for surgeons to seamlessly control the robot. In the tilt camera mode, which is activated by a different gaze gesture, the system allows the camera to rotate the view around the laparoscope’s longitudinal axis.

On initialising the activate camera control mode, the distance from the screen to the user’s eyes is calculated and stored as the original distance. During this control mode, the camera will pan in the vector direction of the PoR from the screen centre with a speed in accordance to Fig. 5 as long as the user’s head is within a region of ±5 cm from the original distance. In the panning mode, the screen area is separated into three radial speed regions as shown in Fig. 5(a). Gazing within each region moves the camera at a different velocity accordingly. If the surgeon’s PoR is within the central screen region, the camera remains stationary allowing for a stable view whilst performing tasks. If the surgeon’s gaze falls within the medium and fast regions, the camera moves at a velocity of 16.9 mms⁻¹ and 23.3 mms⁻¹ respectively. These three speed regions were introduced to enable user-friendly control of the camera whilst maintaining a predefined maximum velocity that would be safe for the patient. Thus, even if the surgeon performs a glance toward the edge of the screen, the camera still follows the predefined safe camera speed. In order to zoom in, the user is required to lean their head forward by more than 5 cm from the original distance position while the activate camera mode is active. This control scheme is illustrated in Fig. 5(b). While zooming in, the surgeon can also direct the camera to simultaneous perform panning motions. Conversely, the user can zoom out during the activate camera mode by leaning their head back 5 cm from the original distance position, as shown in Fig. 5(c). Panning via gaze direction is disabled while zooming outwards.

If the tilt camera gaze gesture is activated, the original left and right eye positions are recorded and the horizon going through the left and right eye is calculated and stored as the base line. The horizon is continuously tracked by updating the position of the left and right eye. The angle formed by the current horizon and the base line is used to control the camera tilt. In order to tilt the camera left or right, the user is instructed to tilt their head by more than ±15° to the left or right respectively, as illustrated in Fig. 5(d). The robot performs incremental rotation steps with an angular velocity of 5° s⁻¹ in the respective tilt direction. In order to maintain a tilt in one direction, the user simply maintains their head position i.e. a tilt of the head greater than 15° will maintain the camera tilt motion in that given direction. Whilst using the system, the user can see their PoR represented on the screen as a white moving dot, which can optionally be deactivated.

Guidance text is overlaid onto the camera view, and the camera can also be stopped by fixing the stop camera text at the bottom left hand corner of the screen (Fig. 5(e)). The stop camera command is identified by detecting dwell-time fixations of at least 750 ms in
that region of the screen. In order to address the potential uncertainties when the eye tracker loses tracking of the user’s eyes, a safety mechanism is introduced where the robotic system immediately stops under lost gaze tracking circumstances. On re-detection of the user’s gaze, the robotic laparoscope resumes with the same control mode as before the tracking was lost.

The control parameters for the three different camera velocities are shown in Table 1. These motion parameters are based on the normalized Euclidean distance \( r \) of the gaze from the centre of the screen as described above. The specific values of these parameters have been chosen for the pan and zoom control modalities according to clinical requirements, and to avoid unexpected motions in joint space. The zooming out motion has the following constant parameters \( L_d = 0.0 \) m, \( L_z = 0.005 \) m and \( t_f = 0.5 \) s.

2.2. Online calibration of gaze tracking

2.2.1. Motivation

Conventional remote gaze trackers require an explicit offline calibration procedure to map the optical axis (OA) (i.e. the straight line that passes through both the pupil and cornea centre) of the user’s eye to account for their visual axis (VA) (i.e. the PoR, or the line joining the cornea and the fovea centre) (Hansen and Qiang, 2010). This process typically requires the user to fixate on a moving spot presented on the screen. During this procedure a set of the user’s uncalibrated PoR coordinates at a number of pre-determined locations on the screen, also known as calibration points, are recorded. The PoR of the user is then corrected using a mapping function which relates the captured PoR coordinates to their respective screen coordinates. A known problem of offline calibra-

| Region                  | Motion Parameters | Condition |
|-------------------------|-------------------|-----------|
| Pan Only                | \( L_d = 0.0 \)  | \( L_d = 0.0, L_z = 0.0 \)  |
| Pan and Zoom-in Simultaneously | \( L_d = 5.5 \) | \( L_d = 6.0, L_z = 5.0 \)  |
| Fast                    | \( L_d = 7.0 \)  | \( L_d = 8.0, L_z = 5.0 \)  |
|                         | \( t_f = 0.010 \)| \( r < r_2 = 0.18 \)  |
|                         | \( t_f = 0.325 \)| \( r_1 < r < r_2 = 0.38 \) |
|                         | \( t_f = 0.300 \)| \( r \geq r_2 = 0.38 \)  |
tion reported in numerous cases is that the calibration can drift over time (Hornof and Halverson, 2002; Nystöm et al., 2013), thus affecting the accuracy and precision of the estimated PoR and potentially requiring the surgeon to recalibrate during surgery. This deterioration of the offline calibration is typically associated with naturally occurring changes in user’s postures or head positions. A quantitative study of drift during laparoscopic surgery is detailed in Appendix B. Given that the tilt and zoom control modes require explicit head movements, the need for online calibration is even more critical in the presented work. Previous gaze tracker systems have used as little as one calibration point with an accuracy of 1° of visual angle. However, these systems utilise multiple cameras and or light sources (Hansen and Qiang, 2010), and still require an explicit offline calibration that is susceptible to calibration drift. Calibration drift would not only lead to poor user-experience, but also raise safety concerns for use in the operating theatre.

To overcome these problems, an implicit online calibration process that progressively adapts to the user’s changing gaze is introduced. Since the proposed online calibration process replaces the conventional offline calibration, the surgeon will be able to use the robotic laparoscope system immediately. Furthermore, the adaptive nature of the algorithm overcomes the calibration drift as it updates with continued use, thus allowing the surgeon to use the gaze contingent system for longer periods without recalibrating during an operation. The online calibration algorithm takes advantage of the pre-learnt gaze gesture information to extract relevant PoR coordinates in an ongoing manner to form and update the mapping function. The proposed online calibration algorithm can be applied to any remote gaze tracker system as long as it possesses user-interactive elements with known positions on the screen, such as menu navigation with the user’s PoR, eye typing, or others such as an automatic scroll mechanism during reading. Once user interaction is recognised, calibration points can be captured and used to remap the user’s PoR. Unlike in Chen and Ji (2015), no assumptions are made on the content of the camera image for the online calibration to function. The presented online approach integrates seamlessly within the gaze gesture framework by taking advantage of the same probabilistic approach used to identify the gaze gestures.

2.2.2. Online gaze calibration design

The gaze gestures require the user to look at the centre and one of the bottom corners of the screen. By extracting the PoR coordinates at these instances, the online calibration process uses these coordinates to populate the subject-specific calibration mapping on the fly. The assumption behind the online calibration is that the user is looking at specific areas located at the center and corners of the screen to perform the gestures. This assumption is made valid because the corner locations are made explicit with text describing the control mode, and because users are trained to use the gestures beforehand. This is further restricted by requiring gestures to be in certain quadrants of the screen, and forming a specific pattern with a particular orientation, i.e. of the form of a gaze gesture that was used to train the gaze gesture model. As shown in Fig. 6, the online calibration process first applies a median filter to the stream of unmapped PoR coordinates from the gaze tracker to reduce noise. Potential gaze gestures are then extracted in the same manner as in the gaze gesture recognition process in Fig. 3. At this stage the gaze gesture sequence consists of a series of unmapped and therefore inaccurate PoR coordinates. To determine whether a potential gaze gesture is not a false positive, principal component analysis is applied to the segmented potential gaze gesture. The majority of the trajectory’s information is contained in the first and second principal components (PC1 and PC2 respectively), and in the form of an elongated diagonal. As such, gaze gestures with a PC1/PC2 ratio below a threshold of 5 are counted as false positives and filtered out. The angle of PC1 indicates the quadrant location of the potential gaze gesture. It is then possible to distinguish whether the segmented gaze gesture is associated with activate camera (HMM1) or tilt camera (HMM2). If PC1 lies on the fourth quadrant, then it is related to the activate camera mode, while if it lies on the third quadrant then it is associated to the tilt camera mode. The absolute positions of the extracted PoR coordinates centroids from the gaze gesture are also stored to rule out inadequate gestures towards the upper left/right corners of the image.

If the principal component criteria are met, the series of unmapped PoR coordinates during two fixation points, i.e. the coordinates associated with the calibration point at the screen centre and the coordinates associated with one of the calibration points at the bottom corner, are extracted from the gaze gesture. Only the PoR coordinates during the period when the velocity is below 300° s⁻¹ are stored.

The stored coordinates are subsequently filtered by computing the centroid and standard deviation of each PoR coordinate within their respective buffers. The final PoR centroid to be used in the mapping function is recomputed excluding any coordinates that fall outside one standard deviation from the initial computed centroid. These centroids are then used in the calibration mapping to map the OA to the VA. The calibration mapping incorporated in the algorithm is a thin plate spline (TPS) based radial basis function (RBF) mapping. TPS, a special polyharmonic spline, was chosen for the gaze calibration mapping due to its elegant characteristics to interpolate surfaces over scattered data. The TPS was first introduced by Duchon (1977), and has previously been used for various computer vision and biological data such as in image registration (Bookstein, 1989). Commercial eye trackers such as the one used in this paper are prone to user-specific errors in gaze tracking. As such, these products typically require an additional calibration procedure to be performed by users prior to working with the eye tracker. The TPS method presented in this work effectively replaces the commercially provided additional calibration procedure with one of our own. Additionally, the proposed calibration procedure is explicit rather than implicit, thus making it much more user friendly. Further details on how the TPS is implemented as a mapping function can be found in Appendix A.

A minimum of three calibration points will be needed to compute the TPS mapping. However, the gaze contingent laparoscope system utilises two gaze gestures, giving room to obtain between two and three calibration points only. Therefore, prior to using the calibration points for the online mapping, the final calibration points associated with the centre, bottom left and right corners of the screen are extrapolated to increase the number of calibration points to five points. Note that on initialisation of the system, there can only be two calibration points obtained from the first gaze gesture received from the user, i.e. either a activate camera or tilt camera gaze gesture. Therefore, symmetry along both the vertical and horizontal eye rotation is assumed and the two calibration points are extrapolated to five calibration points when the first gaze gesture has been successfully performed. This scenario is illustrated in Fig. 8(a-b). When both gaze gestures have been performed, the three calibration points are used to extrapolate to five calibration points as shown in Fig. 8(c). A conventional offline calibration procedure uses anything between five to nine calibration points to establish a gaze mapping function for accurate PoR estimation.

Once five calibration points have been extrapolated, the relevant mapping parameters can be obtained by solving a linear system of equations together with the calibration screen coordinates shown in Fig. 8(d).

Once a calibration mapping is formed, the previously segmented gaze gesture is remapped via the calibration mapping and tested against the two gaze gesture HMM models. If the gaze gesture returns an inference value above either of the two HMM
thresholds, the extracted pupil coordinates are deemed accurate and are kept in respective circular buffers as calibration points, otherwise they are discarded. As the user continues to use the gaze contingent laparoscope system and inputs gaze gestures, the stored pupil coordinate points can be used to build a more robust gaze tracking calibration mapping, whilst also accounting for any calibration drift as the user moves around. The overall online calibration algorithm integrates closely with the gaze gesture recognition algorithm, enabling the surgeon to seamlessly start using the robotic laparoscope system without having to perform an offline gaze tracker calibration. The complete operative workflow is illustrated in Fig. 7, where a surgical resident performs a lesion removal task on an upper gastrointestinal phantom.

3. Results

3.1. Gaze contingent laparoscope user study

In order to assess the accuracy of the gaze gesture recognition and examine the usability of the proposed system by comparing it to other methods, subjects were asked to perform the same task using three different camera control schemes:

1. The proposed gaze gesture control: gesture-based mode activation and camera control through PoR and head position;
2. Pedal activated control: dual-switch foot-pedal mode activation and camera control through PoR and head position;
3. Camera assistant control: a camera assistant follows the verbal instructions of the participant and navigates the camera.

In the foot-pedal control mode the activate camera and tilt camera modes are activated via the left and right pedals respectively. The pedals need to be kept pressed to maintain the chosen control mode, and the camera movement is stopped when the user releases the foot pedal. The laparoscope is navigated in exactly the same manner as shown in Fig. 5(a-d).

3.1.1. Experimental setup

The experimental setup is identical to the one described in Section 2.1.1 (Fig. 1). The HMM gaze recognition process and the robot control process were implemented in C++. The gaze gesture recognition process ran at 33.3 Hz whilst the robot control process updated at 200 Hz. Experimental data which consisted of subject PoR, gaze gestures, and camera-view feed were recorded at a rate of 33.3 Hz. The surgical instrument tip trajectories were recorded with the Polaris at 17 Hz. During the camera assistant control mode the laparoscope tip position was tracked with a Polaris marker. In other modes it was instead obtained from the robot forward kinematics. Instrument trajectory tracking was undertaken as peer-reviewed literature has shown that instrument trajectory path length correlate to the level of surgical performance (Hove et al., 2010).

3.1.2. Participants

In this usability study, seventeen surgical residents with a postgraduate year between 3 – 7 (PGY3-7, male = 16, female = 1) were recruited. The mean laparoscopic experience was 676 (±293) cases. All participants were trained to use the gaze gesture and pedal activated systems on an abstract navigation task before starting the study. This training was performed to prevent potential learning effects when performing the subsequent phantom based task. The abstract training task required the subject to navigate the laparoscope system inside a conventional box trainer to locate numbers in ascending order. The numbers of varying font sizes were placed randomly on a 4 x 5 grid in order to require the user to both pan and zoom during the training. Subject training was halted when: a minimum baseline proficiency task completion time was met; when they showed no further improvement in completion time; and, when they could reproduce a similar completion time consecutively on three occasions. A second training task was used for the tilt control modality of the camera. Subjects were asked to re-align three operative scenes to a conventional anatomical orientation. The task involved tilting the camera left, right, then left by 15°, 65° and 35° respectively. Once a scene was correctly re-aligned, the next scene was presented to the participant.

3.1.3. Tasks

The task involved subjects identifying and removing a set number of randomly placed lesions on an upper gastrointestinal phantom. The task was a simulated upper gastrointestinal staging laparoscopy and the phantom was placed in a laparoscopic box trainer. The nature of the simulated task required subjects to use a bimanual technique, typically with one instrument manipulating and/or retracting tissue, and the other grasping and removing the lesion. The surgeons were allowed and encouraged to physically look at the phantom model before lesions were placed to familiarise themselves with it. This procedure was introduced to minimise the potential confounding factor of learning the phantom model. Participants were asked to perform the lesion removing task twice, for each of the three camera control modes mentioned previously in Section 2.2, namely, via i) gaze gesture activation, ii) foot-pedal activation, and iii) verbal communication with a human camera assistant. Thus, overall each participant performed the lesion removal task six times in total. To mitigate learning effects on performing the lesion removing task, the sequence in which subjects performed the task during the three control modes was randomised. Prior to the user trials, the human camera assistant was given both hands-on and theoretical training over a period of two days on the experimental model by an expert laparoscopic assistant with over 1000 cases performed. The assistant was recalled a week later to confirm retention and proficiency on the experimental model before the study commenced. The assistant was kept constant for all participants.

3.1.4. Performance metrics

Eye tracking data for each participant was recorded during the gaze gesture control mode to assess gaze gesture usability. A high performance gaze gesture recognition algorithm plays a critical role in the usability of the gaze contingent system. In order to assess it, post-hoc observation of the recorded videos of the laparoscopic camera view, was conducted by two independent observers. The PoR data was overlaid post-hoc on all respective camera-view videos. Text was overlaid to signal when the gesture was recognised by the gaze gesture recognition system, making it easier for the observer to count the true positive, false positive and false negative gaze gestures. Observations were performed on all subject videos consisting of a total of 34 videos (17 subjects each of who performed two repetitions of the gaze gesture trial). The two observers viewed video sequences in the same order where their observations were compared for inter-rater reliability using the intraclass correlation coefficient (ICC) (Koch, 1982). The observers recorded the occurrences of true positive (i.e. correctly identified), false positive (i.e. incorrectly identified) and false negative (i.e. incorrectly rejected) gaze gestures in an ordinal manner. To simplify the counting process of the true positive, false positive and false negative gaze gestures, eye-tracking data and the camera-view videos were saved and analysed after the experiments were completed. True negative (i.e. correctly rejected) gaze gestures were obtain by filtering the recorded gaze data and counting the rejected potential gaze gestures during the trials. The recall
is obtained by:
\[
\frac{\sum_{\text{TruePositiveGestures}}}{\sum_{\text{TruePositiveGestures}} + \sum_{\text{FalseNegativeGestures}}}
\]
and the false positive rate is obtained by:
\[
\frac{\sum_{\text{FalsePositiveGestures}}}{\sum_{\text{FalsePositiveGestures}} + \sum_{\text{TrueNegativeGestures}}}
\]

The discriminability index \( d' \) was also computed to evaluate the robustness of the HMM gaze gesture recognition algorithm:
\[
d' = z(\text{Recall}) - z(\text{FalsePositiveRate}),
\]
where \( z(x) \) is the \( z \)-transform. The theoretical limit of \( d' \) is 6.93 and values of \( d' \) of at least 3 were considered acceptable. In addition, the system usability during each laparoscopy control mode is assessed quantitatively through the use of the following performance metrics:

- Task completion time measured in seconds.
- Camera path length measured over a single trial in centimetres - to assess camera control efficiency and usability of the system.
- Camera workspace measured in centimetre cube - to assess whether the group of participants were able to move the laparoscope system over a comparable workspace to that of a camera assistant.
- Instrument path length measured over a single trial in centimetres - to assess the ergonomics of the system.
- Time-normalised instrument path length measured in centimetres per second - to assess the ergonomics of the system. This metric is obtained by dividing the camera path length measured over a single trial by the respective task completion time to obtain a path length independent of task completion time.
- National Aeronautical Space Agency - Task Load Index (NASA-TLX) questionnaire (Hart and Staveland, 1988) - subjects completed a questionnaire after each lesion removal task. This subjective questionnaire is validated and comprises of six variably weighted parameters that contribute to task workload (Zheng et al., 2012a).

The average number of gestures assessed by the two raters was 313 for the activate camera gesture (an average of 18.4 gestures per subject), and 109 for the tilt camera gesture (an average of 6.4 gestures per subject).

### 3.1.5. Gaze contingent laparoscope usability results

The usability performance of the new gaze contingent laparoscope system was based upon the results obtained from the following statistical analysis studies.

1. HMM gaze gesture recall and false positive rate assessment.
2. Comparative analysis of the three different control modalities.
3. Comparative analysis of results from this study against a previously suggested system (Fuji et al., 2013).

At the end, subjective feedback collected from the users regarding the user experience of using the system is presented. For all the statistical analysis, normality tests (Lilliefors test) were initially performed. Normality tests at the 5% significance level revealed the non-parametric nature of the obtained experimental data. Study 2 was a within-subject design; therefore a Wilcoxon signed-rank test was conducted for non-parametric statistical comparison between variables. Results are represented as medians with interquartile ranges (IQR) in brackets, along with respective z and p-values. A \( p\)-value < .05 was considered significant. Results with significant differences are indicated with an asterisk “*” mark in all tables.

In contrast, Study 3 was a between-subject design and therefore Mann Whitney U tests were conducted for non-parametric continuous variables between modalities. Prior to conducting Mann Whitney U tests, Brown-Forsythe F-tests were conducted to check the equal variance assumption. Mann Whitney U tests were then performed only on the data which passed the equal variance criteria. For large samples (U is approximately normally distributed and therefore in this case, standard normal distribution z values are reported.

#### 3.1.6. Hidden Markov model gaze gesture performance assessment

The recall and false positive rate results are shown in Table 2. The overall average recall for the HMM based gaze gestures is 96.48% with an average false positive rate of 1.17%. The discriminability index \( d' \) was 3.706 for the activate camera gaze gesture and 4.714 for the tilt camera gaze, which showed good robustness to visual search behaviour noise. The ICC from observing gaze gestures of both trials for all 17 subjects (34 videos in total) resulted in 0.957 and 0.912 for the activate camera and tilt camera gaze gestures respectively. This result shows strong inter-rater agreement, as coefficient values of greater than 0.8 are typically considered strong agreement. This was not surprising given that identification of the three-stroke gaze gestures was straightforward and unambiguous. More importantly, these results demonstrate that gaze gestures provide high recall and a low false positive rate, making the use of HMM based gaze gestures both user-friendly and safe.

#### 3.1.7. Comparative analysis of different control modalities

The comparative analysis of the three different control modalities uses the performance metrics from the combined data of both trials (first and second trial) and is shown in Table 3. The experiment was a within-subject design, with all seventeen subjects completing two repetitions of all three control modalities. All seventeen subjects met the baseline proficiency and training requirements to be included in the user performance based quantitative analysis. Three of the subjects wore glasses and four wore contact lenses.

The task completion time over the trials were significantly shorter for the gaze gesture activated system compared to both the camera assistant and the pedal activated control scheme (190.50 s vs. 240.50 s; \( z = -1.992, p = .046 \) and (190.50 s vs. 246.00 s; \( z = -3.351, p < .001 \)) respectively. In contrast to this, the pedal activated control modality showed no statistical difference in the task completion time versus the camera assistant (246.00 s vs. 240.50 s; \( z = 0.530, p = .596 \)). The comparative task completion times are shown in the box plot in Fig. 9. Fig. 9 shows that the overall subject group could complete the same bimanual task faster with the gaze gesture activated control modality than both the pedal activated system and the camera assistant.

| Gesture Type   | Recall | False Positive Rate | ICC  |
|----------------|--------|----------------------|------|
| Activate Camera | 96.48% | 1.17%                | 0.957|
| Tilt Camera    | 92.08% | 0.31%                | 0.912|

### Table 2

Recall and false positive rate for both HMM gaze gestures with respective ICC for inter-rater reliability assessment.

![Fig. 9. Task completion time by control mode aggregating all participants.](image-url)
Table 3

User performance metrics of the assistant, gaze gesture activated, and pedal activated system over both trials with respective Wilcoxon signed-rank test results.

| Task         | Can Path Len(cm) | Left Instr Path Len(cm) | Right Instr Path Len(cm) | Time Norm Left Instr Path Len(cm.s$^{-1}$) | Time Norm Right Instr Path Len(cm.s$^{-1}$) | NASA-TLX score |
|--------------|------------------|-------------------------|--------------------------|--------------------------------------------|--------------------------------------------|----------------|
| (1) Assistant| 240.50           | 449.46                  | 1155.71                  | 1250.16                                    | 4.75                                       | 6.02           |
|              | [118.75]         | [291.12]                | [634.70]                 | [699.20]                                   | [1.00]                                     | [1.89]         |
| (2) Gaze Gesture | 190.50           | 975.8                   | 730.81                   | 725.93                                     | 4.45                                       | 5.16           |
|              | [67.50]          | [46.02]                 | [328.70]                 | [336.29]                                   | [1.17]                                     | [1.89]         |
| (3) Pedal    | 246.00           | 104.81                  | 868.26                   | 896.05                                     | 4.15                                       | 4.71           |
|              | [159.50]         | [61.13]                 | [513.18]                 | [449.48]                                   | [1.59]                                     | [1.29]         |
| Wilcoxon Signed-Rank Test | Task | Can Path Len | Left Instr Path Len | Right Instr Path Len | Time Norm Left Instr Path Len(cm.s$^{-1}$) | Time Norm Right Instr Path Len(cm.s$^{-1}$) | NASA-TLX score |
| (2) vs. (1)  | $z = 1.991$      | $p = 0.046^*$           | $z = 4.076$              | $z = 4.847$                                | $z = 3.479$                                | $z = 4.026$    |
|              | $p < 0.001^*$    | $p < 0.001^*$           | $p < 0.001^*$            | $p < 0.001^*$                              | $p < 0.001^*$                              | $p < 0.001^*$  |
| (3) vs. (1)  | $z = 0.530$      | $p = 0.596$             | $z = 3.500$              | $z = 3.548$                                | $z = 3.377$                                | $z = 1.787$    |
|              | $p < 0.001^*$    | $p < 0.001^*$           | $p < 0.001^*$            | $p < 0.001^*$                              | $p < 0.001^*$                              | $p = 0.075$    |

Using the gaze gesture activated system, the surgeon can maintain their visual attention on the laparoscopy camera view without distractions. This makes a significant difference, shown in Fig. 9, as switching control mode completely changes the robotic endoscope behavior. Having to rely on an external component (e.g. an assistant or foot-pedal) to change the robot behavior forces an interruption of the workflow, whereas the gaze gestures allow the surgeon to remain focused on the same medium used to control the robot, without the risk of unintended device motions during the switch.

Significantly shorter camera path lengths were observed for the gaze gesture and pedal activation modalities compared to the camera assistant (97.58 cm vs. 449.46 cm; $z = 5.086$, $p < 0.001$) and (104.81 cm vs. 449.46 cm; $z = 5.086$, $p < 0.001$) respectively. Note that the camera path length showed no statistical difference between the gaze gesture activated and pedal activated control modes (97.58 cm vs. 104.81 cm; $z = 1.000$, $p = 0.317$). Illustration of the group’s comparative camera path lengths is shown in the box plot in Fig. 10. The gaze gesture activated control mode resulted in a significantly shorter task completion time but not in a significant shorter camera path length when compared to the pedal mode. One plausible reason for this result could be the better ergonomics of not having to depress an external pedal device whilst performing a bimanual instrument task. The pressing of a pedal can change the surgeon's posture and balance, thus adversely affecting the ergonomics during the task.

In order to assess whether the gaze contingent system is able to cover a workspace volume comparable to that of a human camera assistant, the camera tip trajectories from the group of surgeons were combined into one point cloud for each control modality. The point clouds were used to obtain a surface mesh via Delaunay triangulation and the overall volume occupied by the camera tip workspace was then computed using the convex hull algorithm. As can be seen from the illustration in Fig. 11, all three camera control methods show a similar workspace volume of 2558.47 cm$^3$, 2556.40 cm$^3$ and 2159.60 cm$^3$ for the camera assistant, gaze gesture activated and pedal activated control schemes respectively.

The instrument path lengths (both left and right) were significantly shorter during the use of the gaze gesture activated control scheme compared to the camera assistant with path lengths at (730.81 cm vs. 1155.71 cm; $z = -4.078$, $p < 0.001$) and (725.93 cm vs. 1250.16 cm; $z = -4.847$, $p < 0.001$) for the left and right instruments respectively. Similarly, the instrument path lengths were also significantly shorter when using the pedal activated control scheme when compared against the camera assistant with left and right path lengths of (868.26 cm vs. 1155.71 cm; $z = -2.094$, $p = 0.036$) and (896.05 cm vs. 1250.16 cm; $z = -3.616$, $p < 0.001$) respectively. Importantly, the instrument path lengths recorded during the gaze gesture activated control scheme also resulted in statistically shorter path lengths when compared against the pedal activated camera control scheme with (730.81 cm vs. 868.26 cm; $z = -3.548$, $p < 0.001$) and (725.93 cm vs. 896.05 cm; $z = -3.377$, $p < 0.001$) for left and right instruments respectively. The group's instrument path lengths during each control scheme are shown in Fig. 12. The shorter instrument path lengths could have been due to the faster task completion times. In order to reduce the time dependencies on the instrument path length, the time normalised instrument path lengths are compared next.

The time normalised instrument path length during the gaze gesture activated control scheme was significantly shorter for both the left and right instrument (4.45 cm.s$^{-1}$ vs. 4.75 cm.s$^{-1}$; $z = -3.479$, $p < 0.001$) and (5.16 cm.s$^{-1}$ vs. 6.02 cm.s$^{-1}$; $z = -4.026$, $p < 0.001$) respectively when compared against those obtained during using a camera assistant. Time normalised instrument path lengths during the pedal activation control scheme were also significantly shorter for both left and right instruments compared to the camera assistant trials (4.15 cm.s$^{-1}$ vs. 4.75 cm.s$^{-1}$; $z = -3.351$, $p < 0.001$) and (4.71 cm.s$^{-1}$ vs. 6.02 cm.s$^{-1}$; $z = -4.727$, $p < 0.001$) respectively. There was no statistical difference in the time normalised instrument path length between.
the gaze gesture activated and pedal activated control modality for either instrument. Given the within-subject design, and that participants were experienced surgeons, the shorter time normalised instrument path length reflects the improved usability of the gaze contingent control modes. A shorter instrument path length, which is a reflection of efficient instrument movement, has previously been associated with better surgical performance in the clinical setting (Aggarwal et al., 2007; Van Sickle et al., 2005). Since the group of surgeons participating in the study remained the same, any changes in the instrument path length can be inferred as an indirect measure of usability.

Each control scheme was also assessed for its contribution to the cognitive workload of the participant through the NASA-TLX questionnaire. A desired aspect of new technology introduced in the operating theatre is one which does not add to the cognitive burden of the surgeon. No statistically significant difference can be observed in the NASA-TLX score outcome for the gaze gesture activated control scheme relative to the camera control using a camera assistant (40.00 vs. 32.67; \( z = 1.188, p = .235 \)). However, the foot-pedal activation method resulted in significantly higher NASA-TLX scores compared to the camera assistant mode (45.34 vs. 32.67; \( z = 2.137, p = .033 \)). The change in the user's balance and posture when required to use an additional limb to depress a pedal and activate the camera might be the cause of the disparity in the NASA-TLX scores. The overall group NASA-TLX scores are shown in Fig. 13.

3.1.8. Comparative analysis to an alternative gaze contingent system

The final usability analysis involves a comparison of the system presented in this article to that presented by Fujii et al. (2013). Both systems use the same Kuka LWR arm and two gaze gestures to control the camera. Furthermore, the same panning and zooming speeds are used in both systems. The main difference between the two systems is the separation of the pan and zoom control in the previous work. In order to switch from panning the camera to zooming, the user would have to stop the camera and then perform a gaze gesture to switch to zoom control. In contrast, the system UI presented in this article combines the pan and zoom control into one activate camera control mode, where the user can switch between the panning and zooming by moving their head forward or backward. In addition, the new system enables an extra tilt camera control to rotate the camera view along the laparoscope's longitudinal axis.

The work by Fujii et al. (2013) had a subject group size of eleven participants with laparoscopic experience of 536 (\( \pm 315 \)) cases. The experience of the participating group of surgeons is comparable to the experience of the group of surgeons recruited for the user trials presented in this article, which is seventeen subjects with laparoscopic experience of 676 (\( \pm 293 \)) cases. A similar task to that in Fujii et al. (2013) was completed by the subjects. To analyse the between-subject designed user study, Brown–Forsythe F-tests were performed to check comparable variance of the comparision group data and subsequently Mann Whitney U tests were performed comparing the two systems. A summary of these results are presented in Table 4.

From the Brown–Forsythe F-test, the only between-subject group pair that did not meet the equal variance criteria was the task time results obtained from our proposed gaze gesture modality and the task time from the gaze gesture modality of Fujii et al. (2013). The failure to show equal variance of these two grouped task time data implies that the task time obtained during the use of the proposed gaze gesture modality had a significantly different variance to the one obtained during the Fujii et al. method (190.50 [67.50]s vs. 281.00 [172.00]s). Thus, this F-test shows improved consistency and speed in achieved task times by participating surgeons, when using the proposed gaze gesture control scheme. In contrast, as shown by the Mann Whitney U tests, the pedal activated control method showed no significant difference in task completion times. This result could indicate that our system, which enables quick switching between panning and zooming, is more ergonomic when the gaze gestures are used to activate the camera but not necessarily when the user is required to depress a foot-pedal. The subtle change in the user's balance and posture when having to use another limb to depress the pedal and activate the camera might be the cause of the disparity.

The camera path lengths resulted in no statistical difference for both the gaze gesture and the pedal activated control schemes. This result is not surprising as the system's parameters for panning and zooming were kept identical. Therefore, the significantly different variance statistic in the task time resulting from subject trials during the use of our proposed gaze gesture activated control scheme is likely due to the better system ergonomics. The NASA-TLX showed no significant change between the two gaze gesture activated control schemes. Interestingly, the pedal-activated control scheme that we propose shows a significant increase in the NASA-TLX score when compared to the pedal system by Fujii et al. (2013) resulting in (45.34 vs. 27.00; \( z = -2.383, n_1 = 34, n_2 = 20, p = .017 \). This increase in cognitive burden for the cohort of surgeons may have been caused by the restricted posture during zooming in or out. Since the foot-pedal needs to maintain depressed during camera control, the limb used for the
foot-pedal is essentially in a fixed position. Since the new system requires moving their head forwards or backwards during zooming in or out, the combined UI may have been awkward at times for the surgeon.

3.1.9. Subjective feedback
Surgeons who participated in this study were also asked for subjective feedback. Most of the feedback was positive, including comments such as that the panning control was working effectively and provided the advantage of maintaining a steady camera view and horizon compared to a camera assistant. Some surgeons expressed their preference for the gaze gesture activated system over the pedal system, with the opinion that it is easier to learn and to use, while the addition of the foot pedal to the gaze control increases their cognitive demand when moving the camera. On the other hand, some surgeons expressed that although the system is intuitive to use, they would prefer to use a human as that is what they are accustomed to. However, these surgeons also expressed that they could see the benefit of the system especially for long operations which would be unpleasant for an assistant. Some surgeons felt that the stop camera location, which was at the bottom left corner of the screen conflicted with the usability as it causes the camera to move slightly while they fixate at the corner for 750 ms. Other feedback included some personalised preferences including a desire to have faster pan or zoom speed.

3.2. Online gaze calibration performance study
The aim of this study was to assess whether the online calibration algorithm could calibrate “on the fly” with a range of different subjects and maintain a high level of accuracy and precision over time. Furthermore, the study compares the online calibration algorithm performance to when the gaze tracker is not calibrated, and when an offline 5 and 9 point calibration procedure is conducted. All gaze gestures performed were recorded for offline analysis to quantify the recall and false positive rate. As the purpose of this study was to assess the accuracy of the online calibration method through periodic checks, it was performed independently from the study presented in Section 3.1 which was meant to simulate an uninterrupted surgical scenario as closely as possible.

3.2.2. Participants
Twenty-five subjects participated in the within-subject user study to assess the performance of the online calibration algorithm independently (male = 20, female = 5). All participants were trained to use the gaze gesture UI before starting the study.

3.2.3. Tasks
Each participant was required to successfully perform both gaze gestures ten times as training. Post training, each participant was asked to perform a calibration performance task under the following eye tracker calibration conditions: i) no calibration, ii) five point offline calibration, iii) nine point offline calibration, iv) after one gaze gesture (online calibration), v) after two gaze gestures (both activate camera and tilt camera gesture (online calibration), vi) after five gaze gestures (online calibration), vii) after ten gaze gestures (online calibration). The performance task involved the participant observing one by one, nine evenly distributed white dots displayed on a screen. During this task, the participants’ gaze was recorded for offline analysis.

Each trial was carried out over twenty minutes. The participant performed one gaze gesture, then the performance task, then was asked to take a five minute break by moving away from the desk. Subsequently, the same procedure was also repeated after performing two, five and ten gaze gestures in the same session. The gaze gesture count is accumulated to emulate the subject performing the calibration on the go. The study was executed in this manner to understand the online calibration’s longitudinal performance.

3.2.4. Performance metrics
In order to measure the accuracy of each calibration method the angular divergence $\theta$ of the user’s fixation point from the nine reference points is computed:

$$\theta = 2 \times \tan^{-1}\left(\frac{S}{2D}\right).$$

(8)

where $D$ represents the distance in centimetres from the subject’s eye to the gaze tracker screen and $S$ represents the distance offset in centimetres from a reference point as illustrated in Fig. 14.

The overall accuracy of the calibration is defined as,

$$\theta_{\text{overall}} = \frac{1}{n} \times \sum_{i=1}^{n} \theta_i,$$

(9)

where $\theta_i$ is the angular divergence of the recorded fixations from each of the reference targets and $n$ is the number of reference tar-
gets. In addition, the precision of the calibration is defined as,

$$\theta(i)_{\text{precision}} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} \theta_j^2}. \quad (10)$$

The overall precision is obtained by taking the average between the nine precision measurements i.e.,

$$\theta_{\text{AvgPrecision}} = \frac{1}{n} \sum_{i=1}^{n} \theta(i)_{\text{precision}}. \quad (11)$$

The gaze gesture recall during the online calibration is also assessed to check if the online calibration adversely affects the gaze gesture recognition algorithm’s performance. The gaze gesture’s recall, false positive rate and the discriminability index $d’$, are quantified as in Section 3.1 by post-hoc observation of the recorded camera-view videos by two independent observers. The observers viewed the video sequences in the same order where their observations were compared for inter-rater reliability using the ICC. The average number of gestures assessed by the two raters was 136 for the activate camera gaze gesture (an average of 5.44 gestures per subject) and 129 for the tilt camera gaze gesture (an average of 5.16 gestures per subject).

3.2.5. Algorithm performance analysis and results

The performance of the online calibration process is based upon the results obtained from the following studies:

1. Comparative analysis of the online calibration against the offline calibration for accuracy and precision.
2. HMM gaze gesture recall and false positive rate assessment.

For all statistical analyses, normality tests were performed. Normality tests at the 5% significance level revealed the non-parametric nature of the collected subject gaze data. Study 1 was a within-subject design, therefore a Wilcoxon signed-rank test was conducted for non-parametric statistical comparison between variables. Results are represented as medians with IQR in brackets, along with respective z and p-values. A $p$-value < .05 was considered significant. Results with significant differences are indicated with an asterisk ‘*’ mark in all tables.

3.2.6. Online calibration accuracy and precision assessment

The comparative accuracy and precision performance experiment was a within-subject design, with all twenty-five subjects undertaking the respective calibration procedures and the performance recorded for each calibration technique. The comparative accuracy and precision performance of the online gaze tracker calibration algorithm compared to having no calibration and an offline calibration is summarised in Tables 5–7. From these tables, it is observable that the online calibration has consistent PoR estimation accuracy throughout the trial (after performing one, two, five and ten gaze gestures). Offline calibration methods have previously shown to deteriorate over time in accuracy performance (Nystrom et al., 2013) which would in turn hinder gaze tracking techniques to be applied in the surgical theatre. During the fifteen to twenty minute duration of the trial, the gaze tracker’s calibration accuracy is maintained or even improved, which is a desirable attribute.

Wilcoxon signed-rank tests show that the PoR estimation accuracy of the online calibration improves with more gaze gestures; the online calibration accuracy after two, five and ten gaze gestures shows a statistically significant improvement over that obtained after one gaze gesture with PoR accuracy of (0.89 vs. 0.83); $z = 1.968$, $p = .049$, $(0.89 vs. 0.72); z = 2.616$, $p = .009$ and $(0.89 vs. 0.70); z = 5.153$, $p < .001$ respectively. Furthermore, the PoR accuracy of the online calibration after ten gaze gestures shows statistical improvement over the accuracy obtained after two gaze gestures (0.83 vs. 0.70; $z = 3.102$, $p = .002$), five gaze gestures (0.72 vs. 0.70; $z = 2.317$, $p = .021$), as well as the five point offline (0.70 vs. 0.82; $z = -2.922$, $p = .004$) and nine point offline (0.70 vs. 0.80; $z = -2.505$, p = .012) calibrations. The accuracy improves with more gaze gestures as the calibration mapping stores more information, which is reflected in the significant $p$-values of the Wilcoxon signed-rank test.

In addition, the accuracy of the online calibration after one gaze gesture is high at 0.89 [0.74] and after two gaze gestures, the online calibration is comparable to that of the five point offline calibration (0.83 vs. 0.82; $z = 0.203$, p = .839) or nine point offline calibration (0.83 vs. 0.80; $z = 0.269$, p = .788) respectively. The significance test’s $p$-values confirm this statement, as no significant difference was observed between the online calibration’s accuracy after two gaze gestures to that obtained from a five point offline calibration or a nine point offline calibration. As expected, in the absence of any calibration the accuracy is poor (3.54 [2.50]). This result is also consistent with the result of the statistical comparison tests against the online and offline calibration methods, where the accuracy improves significantly after the gaze tracker has been calibrated with any offline or online calibration method.

Previous literature has highlighted drifting of the precision of offline calibration methods with time (Nystrom et al., 2013). In this study, we have shown from the statistical comparison tests that the online calibration algorithm is able to consistently achieve statistically indifferent precision to those of offline calibration techniques, with the added advantage that the online calibration algorithm maintains its precision through prolonged usage. The PoR estimation accuracy with respect to the location of each reference point for the group of 25 participants during different calibration methods are illustrated in Fig. 15. The distance of the lines represent the accuracy error from each reference point. From Fig. 15(a), it is clear that the accuracy of the PoR estimate can vary significantly within the group of subjects when there is no calibration, resulting in inaccurate PoR estimation. Fig. 15(b) and (c) respectively show the PoR estimation accuracy during a nine point offline calibration and an online calibration after ten gaze gestures. The figures illustrate the improvement that can be achieved in the PoR estimation accuracy from having either an offline or online calibration.

3.2.7. Hidden Markov model gaze gesture performance assessment

The last performance analysis of the online gaze tracker calibration algorithm is the evaluation of the recall and false positive rates attained during the use of the algorithm. The results are summarised in Table 8. The overall average recall for the HMM based gaze gestures is 96.81%, with an average false positive rate of 0.60%. The discriminability indices $d’$ for the activate camera and tilt camera gaze gestures were 4.384 and 4.426 respectively, therefore showing good robustness to visual search behaviour noise. The ICC obtained from observing videos of gaze gestures being performed during the online calibration performance assessment for all 25 subjects (a total of 25 videos) resulted in 0.946 and 0.954.
Recall the Online test results (On2) and online Wilcoxon Signed-Rank test results (On1) for the activate camera and tilt camera gaze gesture respectively. The strong agreement between the two observers indicated by the ICC value greater than 0.8 is not surprising given that identification of the three-stroke gaze gestures were straightforward and unambiguous. These results are comparable to those obtained when the gaze tracker is calibrated offline (shown in Table 2), thus demonstrating that the online calibration algorithm does not affect the usability of the gaze gesture activated laparoscope system. Furthermore, the very low false positive rate for the detection of gaze gestures means it is highly unlikely for unintended eye movements, and therefore erroneous gaze gestures, to be used in the calibration.

### 4. Discussion

In this article, we have introduced a gaze contingent robotic laparoscope which allows for pan, tilt and zoom motion capabilities in Cartesian space. Gaze gestures were used to activate the different camera control modes, with a combined panning and forward-backward zooming control mode implemented using only head motions. Validation of the system showed that HMMs are effective in recognising gaze gestures, with mean experimental recall and false positive rate of 96.5% and 1.2% respectively. Results show that user intention can be separated from unintentional eye movements, therefore providing a means to communicate with the robotic laparoscope. Such a method could also be applied to other areas such as helping disabled patients communicate.
A novel online gaze tracker calibration algorithm is also introduced. Experimental results show that the algorithm is able to obtain an accuracy of better than 1° of visual angle with one gaze gesture. Comparable accuracy and precision to that of a conventional offline gaze tracker calibration can be obtained after only two gaze gestures. The results from the calibration performance show that without a calibration procedure the PoR estimation accuracy is poor and can be prone to subject specific variation. During pilot studies it was noticed that a few subjects could have their gaze gestures recognised without calibrating the gaze tracker. However, it would be undesirable from a clinical perspective to have to direct the camera with an inaccurate PoR estimation. The introduction of this online gaze tracker calibration removes the need to perform an offline subject-specific calibration before being able to use the gaze tracker, improving the surgical workflow. Furthermore, online calibration has the added advantage of constantly updating over time, thus avoiding calibration drift problem and resulting in an accurate PoR estimation.

In addition, a comprehensive usability study involving seventeen surgical residents was conducted to assess the new gaze gesture activated robotic laparoscope system. Results demonstrated that once the group of surgeons learnt how to use the system, they were able to perform a surgical navigation task quicker, with a superior camera and instrument efficiency when compared to instructing a camera assistant or when using a pedal activated control scheme. The gaze gestures provide an effective means to convey the surgeon’s desired camera control method, and the seamless switching between panning and zooming in and out by leaning forward or backward are likely to have contributed to the improved user performance. Although pedals are commonly used in the operating theatre today, having to depress a foot-pedal can change the ergonomics of the operation. Analysis of the camera workspace occupied during the user trials demonstrated that the gaze contingent laparoscope system is able to navigate a similar working volume to that of a human camera assistant. The NASA-TLX scores indicated no signs of cognitive burden during the use of the gaze gesture control mode for the group of participants when compared to using the camera assistant or pedal activated control modes. This result therefore suggests that the group of participants did not feel the gaze gesture activated system to be a complex control scheme. However, although the NASA-TLX is a well validated tool for appraisal of subjective workload in general human factors research, the subjective nature of the questionnaire should still be taken into account when interpreting results.

A comparison of the proposed gaze gesture activated method to other systems was also conducted. It has been shown that the technique has faster and more consistent task completion times with lower group variance. The pedal activated control schemes did not show any difference in terms of task completion time, but resulted in significantly higher NASA-TLX scores, indicating that the surgical residents felt a higher cognitive burden whilst using it. This result is perhaps due to the introduction of using head motion to zoom in and out, which could have caused an awkward posture when combined with the need to press a pedal to control the camera.

Overall, the gaze contingent laparoscopic control performed well, allowing the surgeon to rapidly execute a bimanual task without requiring a camera assistant. Furthermore, the gaze gestures were easily learnt and used by the group of participants. However, while the studies presented in this work clearly show the usefulness of gaze contingent data in robotic surgery, some limitations were also highlighted. One such limitation stems from the eye tracking hardware used. In particular, when a surgeon is wearing thick framed spectacles, there is potential for the gaze tracker to have larger PoR estimation errors and experience trouble tracking the user’s gaze. This in turn can impact the gaze gesture detection algorithm. Another hardware limitation is the external workspace of the system. While the robotic arm - laparoscope setup used was sufficient to carry out the experiments comfortably, custom-designed hardware would be able to maximise the workspace available to the surgeon.

Several additional improvements to the system can be made based on the lessons learned from the comprehensive studies performed, as well as the surgeon feedback received. For instance, adjusting the speed regions so that the speed is proportional to the distance of the PoR from the centre of the screen would allow for smoother and more intuitive transitions. Furthermore, surgical feedback included the desire to be able to manually tune the speed of the robot, in order to achieve a pace they are comfortable with. However, due to the nature of the gaze contingent control a compromise must be made between responsiveness and smoothness of motion. Future work will study the impact of using different control schemes for the gaze-contingent laparoscope control. Active constraints can be added to implement safety-boundaries for the robot workspace and machine learning can be incorporated into the online calibration algorithm to enable a more robust calibration algorithm. More gaze gestures can be added, for example towards the top corners of the screen, to create an immersive environment for the surgeon to switch on and off a number of surgical applications intra-operatively, e.g. patient specific visualisations to help localise tumours whilst further improving the online calibration accuracy. Furthermore, reinforcement learning techniques (Wang et al., 2013) can potentially be introduced in the gaze gesture recognition process to improve the personalised recognition performance of the gaze gesture recognition process. Spatially invariant gaze gestures are another area under research to enable head mounted gaze trackers to be used in the surgical theatre. The use of head mounted gaze trackers would offer a larger workspace for the surgeon, as current screen based gaze trackers can only offer consistent tracking accuracy within 1–1.5 m from the gaze tracker. Lastly, assessing the gaze contingent system in multi-disciplinary team environments could also be of interest.

Appendix A. Theoretical formulation of the online gaze calibration

A1. Application of thin plate spline to gaze mapping

Given a vector \( x \in \mathbb{R}^m \), a RBF is any real-valued function \( \phi \) whose value depends only on the distance from a centre vector \( c \in \mathbb{R}^m \), such that \( \phi(x, c) = \phi(\|x - c\|) \), where \( \|\cdot\| \) denotes the Euclidean norm. Sums of RBFs are commonly used to approximate functions. They take the form

\[
    f(x) = \sum_{i=1}^{n} w_i \phi(\|x - c_i\|),
\]

(A.1)

where the approximating function \( f(x) \) is represented as a sum of \( n \) RBFs, each associated with a different centre \( c_i \), and weighted by an appropriate coefficient \( w_i \). This function defines a spatial mapping that maps any location \( x \) in space to a new location \( f(x) \). In order to use the TPS (Bookstein, 1989) as the basis function, the TPS takes the form \( \Phi(r) = r^2 \ln(r) \), where \( r = \|x - c\| \).

To guarantee the unique existence of interpolants with a small variation to Eq. (A.1), lower order polynomials are added to \( f(x) \) along with additional conditions. Specifically, we obtain the following formulation for the TPS:

\[
    f(x) = \sum_{i=1}^{n} w_i \phi(\|x - c_i\|) + a + b^T x.
\]

(A.2)
where \( \mathbf{x} \in \mathbb{R}^m \), \( a \) is a constant, and \( \mathbf{b} \in \mathbb{R}^m \). Eq. (A.2) gives rise to a unique interpolating function \( f(\mathbf{x}) \) using the TPS basis function \( \Phi(\mathbf{r}) = r^2 \ln(r) \), valued at zero for \( r = 0 \). The uniqueness can be guaranteed provided that the centre vectors \( \mathbf{c}_i \) of the basis functions are not collinear, and the following conditions are fulfilled:

\[
\sum_{i=1}^{n} w_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} w_i \mathbf{c}_i = (0, \ldots, 0)^T. \tag{A.3}
\]

where \((0, \ldots, 0)^T\) denotes the zero vector in \( m \) dimensions. Under these conditions, the scalars \( w_i \), \( a \), and the vector \( \mathbf{b} \) can be uniquely solved for.

To use the TPS basis function as the gaze calibration mapping, the unmapped PoR coordinate centroids, \( \mathbf{P} = (x, y) \) are defined as the input, and hence \( m = 2 \). The mapping to the screen coordinates \( \mathbf{f}(\mathbf{P}) \) is represented as:

\[
f(\mathbf{P}) = \sum_{i=1}^{n} w_i \Phi(||\mathbf{P} - \mathbf{c}_i||) + a + \mathbf{b}^T \mathbf{P}. \tag{A.4}
\]

where \( \mathbf{c}_i = (\tilde{x}_i, \tilde{y}_i) \) is the \( i \)th PoR feature vectors obtained during calibration, and \( \mathbf{b} = [a, a'] \). The conditions for the scalars \( w_i \), \( a \), and the vector \( \mathbf{b} \) to be solved uniquely become:

\[
\sum_{i=1}^{n} w_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} w_i \tilde{x}_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} w_i \tilde{y}_i = 0. \tag{A.5}
\]

The mapping in Eq. (A.4) specifies an approximation function \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) in order to describe an eye tracking calibration mapping in \( \mathbb{R}^2 \rightarrow \mathbb{R}^2 \). As such, two TPS functions are used; one for the \( x \)-axis of the screen \( f_x(\mathbf{P}) \) and one for the \( y \)-axis of the screen \( f_y(\mathbf{P}) \). These mapping functions share their \( n \) PoR feature vectors, \( \mathbf{c}_i = (\tilde{x}_i, \tilde{y}_i) \) obtained during calibration to give:

\[
M(\mathbf{P}) = \begin{bmatrix} f_x(\mathbf{P}) \\ f_y(\mathbf{P}) \end{bmatrix} = \sum_{i=1}^{n} \begin{bmatrix} w_i^x \\ w_i^y \end{bmatrix} \Phi(||\mathbf{P} - \mathbf{c}_i||) \\
+ \begin{bmatrix} a^x \\ a'^x \end{bmatrix} + \begin{bmatrix} b^x \\ b'^x \end{bmatrix} \mathbf{P}. \tag{A.6}
\]

Here, \( w_i^x, a^x, b^x \) represent the scalar coefficients and vector for the \( x \)-axis, and \( w_i^y, a'^x, b'^x \) represent the scalar coefficients and vector for the \( y \)-axis. The resulting mapping function \( M(\mathbf{P}) \) now describes the mapping of the PoR coordinates \( \mathbf{P} \) in the eye image plane to the screen pixel coordinate plane. The overall TPS calibration mapping is divided into two parts: the sum of the RBF weighted by the TPS coefficients \( w_i^x \) and \( w_i^y \), which are bounded and asymptotically flat, and an affine part described by the last three terms. Importantly, three coefficients are required to express an affine transform, which means a minimum of three calibration points will be needed to compute the TPS mapping.

### A2. Thin plate spline mapping parameter estimation

Once \( n \) calibration points have been extrapolated, with \( n \geq 3 \), the relevant mapping parameters including \( w_i^x, w_i^y, a^x, a'^x, b^x, \) and \( b'^x \) can be obtained by solving the following linear system:

\[
\begin{bmatrix} S^T \\ 0.3^T \end{bmatrix} = \begin{bmatrix} L \\ C^T \end{bmatrix} \begin{bmatrix} \alpha^x \\ \alpha'^x \\ \mathbf{b}^x \\ \mathbf{b}'^x \end{bmatrix}. \tag{A.7}
\]

where \( S \) is composed of the \( n \) calibration points’ respective screen coordinates:

\[
S = \begin{bmatrix} S_1^T \\ \vdots \\ S_n^T \end{bmatrix} = \begin{bmatrix} S_1^x \\ \vdots \\ S_n^y \end{bmatrix}. \tag{A.8}
\]

as illustrated in Fig. 8(d). \( \mathbf{0}^{3 \times 3} \) and \( \mathbf{0}^{3 \times 2} \) denote the \( 3 \times 3 \) and \( 3 \times 2 \) zero matrices respectively. The calibration points’ respective \( n \) PoR feature coordinates represented by \( \mathbf{c}_i = (\tilde{x}_i, \tilde{y}_i) \) form the array \( \mathbf{C} = (1, \tilde{x}_1, \tilde{y}_1) \), a \( n \times 3 \) matrix. Let \( \mathbf{L} \) be a \( n \times n \) matrix defined as:

\[
\mathbf{L} = \Phi(||\mathbf{c}_j - \mathbf{c}_i||) + \lambda \cdot \mathbf{I} \cdot a^2. \tag{A.9}
\]

where \( \mathbf{I} \) is the \( n \times n \) identity matrix, \( \lambda \) is the regularization coefficient, and:

\[
\alpha = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} ||\mathbf{c}_j - \mathbf{c}_i||. \tag{A.10}
\]

The regularization coefficient \( \lambda \) controls how strictly the model respects the control points. For \( \lambda = 0 \) the TPS interpolation will intersect each control point exactly, at the cost of having a more rapidly varying model between control points. For \( \lambda \rightarrow \infty \), the TPS model approaches a least squares fitted plane. \( \mathbf{t}^\mathbf{x} \) and \( \mathbf{t}' \) are arranged as:

\[
\mathbf{t}^\mathbf{x} = \begin{bmatrix} w_1^x \\ \vdots \\ w_n^x \end{bmatrix}, \quad \mathbf{t}' = \begin{bmatrix} a^x \\ \mathbf{b}^x \end{bmatrix}. \tag{A.11}
\]

Both \( \mathbf{t}^\mathbf{x} \) and \( \mathbf{t}' \) are \( (n + 3) \times 1 \) vectors. By further denoting:

\[
\mathbf{K} = \begin{bmatrix} \mathbf{L} \\ \mathbf{C}^T \\ \mathbf{0}^{3 \times 3} \end{bmatrix}. \tag{A.12}
\]

We can invert the system in Eq. (A.7) to solve for the scalar and vector parameters \( \mathbf{t}^\mathbf{x} \) and \( \mathbf{t}' \) as follows:

\[
\begin{bmatrix} \mathbf{t}^\mathbf{x} \\ \mathbf{t}' \end{bmatrix} = \mathbf{K}^{-1} \begin{bmatrix} S^T \\ 0.3^T \end{bmatrix}. \tag{A.13}
\]

### Appendix B. Quantification of calibration drift

#### B1. Experimental setup

Ten participants were asked to perform a laparoscopic ring transfer task using standard laparoscopic instruments. The same laparoscopic phantom that was used for the paper’s studies was also used for this study. The experiment was divided in 7 steps:

1. An initial offline calibration of the eye tracker.
2. 1 min of laparoscopic ring transfers
3. A second offline calibration
4. 1 min of laparoscopic ring transfers
5. A third offline calibration
6. 1 min of laparoscopic ring transfers, and the participant was asked to turn around and simulate talking to an assistant for 10 s.
7. A final offline calibration.

The participants were not asked to stay particularly still throughout the different steps, to represent natural behavior during surgery. Reaching the specified area inside the phantom to perform the ring transfer was made sufficiently challenging to be able to simulate the motions that might arise from a real procedure.

#### B2. Results

A calibration procedure yields two elements: a sequence of detected gaze points corresponding to the calibration dots displayed on the screen, and a thin plate spline model computed from these points. For a sequence of gaze points issued from a given calibration, Fig. B.1 shows the error computed using:
1. The thin plate spline model generated from that calibration (the calibration self error). Note that the regularization parameter of the thin plate spline model was chosen to insure a smoothly varying model without overfitting.

2. The thin plate spline model generated from the initial calibration

3. The thin plate spline model generated from the previous calibration

Fig. B.2 shows the impact of having the user turn around and talk after having performed the calibration. This done by splitting the data from Fig. B.1 (right) into two categories:

1. The error obtained using the thin plate spline model generated by the first and second calibrations with the gaze points recorded from the second and third calibrations respectively.

2. The error obtained using the thin plate spline model generated by the third calibration with the gaze points recorded from the final calibration.

The results shown in Fig. B.1 highlight the substantial difference between a newly-calibrated system and one where the user has been free to move while performing a challenging task. A newly calibrated system has a non-zero error due to the low-bending model used to prevent overfitting ($\lambda > 0$, see Appendix A). The median error for a newly calibrated system is 6.9 pixels, as opposed to a 34.9 pixel error when using the initial initialization with new gaze data after performing the tasks. Projecting this error on the 24 inch 1920 x 1200 pixel used for the study, this amounts to a 1.8 mm mean error for a newly calibrated system, as opposed to a 9.4 mm error after use. This five times increase in the detection error is easily noticeable in practice, as the mean error after use is nearly a centimeter.

Although not statistically significant relative to the initial calibration error, the median calibration error between successive calibrations is slightly higher, at 44.5 pixels (p-value of .2549). This is not unexpected as motion generated during the task will typically move the user away from their starting position for that task, but these movements would average out over the duration of the entire experiment (i.e. the user is not continuously physically drifting in one direction). Effectively, this result means that the initial position a user had at the beginning of the experiment is closer on average to any given position after a task than the position the user had before starting that particular task.

Finally, Fig. B.2 shows that asking the user to turn around and simulate a quick conversation does not make a statistically significant difference with regards to the calibration drift (p-value of .5824). While the mean, median, and standard deviation of the errors all increase, this result shows that most of the drift already appears during naturally occurring motions.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.media.2017.11.011.

References

Aggarwal, R., Grantcharov, T., Moorbyth, K., Milland, T., Papasavas, P., Dosis, A., et al., 2007. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Ann. Surg. 245 (6), 992–999.

Albu-Schaffer, A., Hafidadi, S., Ott, C., Stemmer, A., Wimböck, T., Hirzinger, G., 2007. The DLR lightweight robot: design and control concepts for robots in human environments. Ind. Robot 34 (5), 376.

Albu-Schaffer, A., Ott, C., Hirzinger, G., 2007. A unified passivity-based control framework for position, torque and impedance control of flexible joint robots. Int. J. Rob. Res. 26 (11), 23.

Bookstein, F.L., 1989. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. 11, 567–585.

Chen, J., Ji, Q., 2015. A probabilistic approach to online eye gaze tracking without explicit personal calibration. IEEE Trans. Image Process. 24 (3), 1076–1086.

Drewes, H., Luca, A.D., Schmidt, A., 2007. Eye-gaze interaction for mobile phones. In: Paper Presented at the Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology.

Duchon, J., 1977. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. 85–100.

Fuji, K., Salerno, A., Srikantharakal, K., Kwok, K.-W., Shetty, K., Yang, G.-Z., 2013. Gaze contingent cartesian control of a robotic arm for laparoscopic surgery. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

Gras, G., Leibrandt, K., Wisanuvej, P., Giataganas, P., Seneci, C.A., Ye, M., Shang, J., Yang, G.-Z., 2017. Implicit gaze-assisted adaptive motion scaling for highly articulated instrument manipulation. In: IEEE International Conference on Robotics and Automation (ICRA).

Gras, G., Yang, G.-Z., 2016. Intention recognition for gaze controlled robotic minimally invasive laser ablation. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2431–2437.

Hansen, D.W., Qiang, J., 2010. In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. 32 (3), 478–500.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

Hornof, A.J., Halverson, T., 2002. Cleaning up systematic error in eye-tracking data by using required fixation locations. Behav. Res. Methods Instrum. Comput. 34 (4), 592–604.

Hove, P.D.v., Tuı́thof, G.J.M., Verdaasdonk, E.G.G., Stassen, L.P.S., Dankelman, J., 2010. Objective assessment of technical surgical skills. Br. J. Surg. 97 (7), 972–987.
