2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Gaze Contingent Cartesian Control of a Robotic Arm for Laparoscopic Surgery

Kenko Fujii, Antonino Salerno, Kumuthan Sriskandarajah, Ka-Wai Kwok, Kunal Shetty, and Guang-Zhong Yang, *IEEE Fellow*

Abstract— **This paper introduces a gaze contingent controlled robotic arm for laparoscopic surgery, based on gaze gestures. The method offers a natural and seamless communication channel between the surgeon and the robotic laparoscope. It offers several advantages in terms of reducing on-screen clutter and efficiently conveying visual intention. The proposed hands-free system enables the surgeon to be part of the robot control feedback loop, allowing user-friendly camera panning and zooming. The proposed platform avoids the limitations of using dwell-time camera control in previous gaze contingent camera control methods. The system represents a true hands-free setup without the need of obtrusive sensors mounted on the surgeon or the use of a foot pedal. Hidden Markov Models (HMMs) were used for real-time gaze gesture recognition. This method was evaluated with a cohort of 11 subjects by using the proposed system to complete a modified upper gastrointestinal staging laparoscopy and biopsy task on a phantom box trainer, with results demonstrating the potential clinical value of the proposed system.**

I. Introduction

In recent years, 'keyhole' (laparoscopic) surgery has established itself as the gold standard for treating a wide variety of surgical conditions. The advantages of the technique include diminished post-operative pain, reduced blood loss, fewer adhesions, shorter hospitalisation and a faster return to normal activities. However, this technique requires a distinctive set of complex surgical skills and advanced training.

To perform laparoscopic surgery, the surgeon typically needs an assistant to orientate and navigate the laparoscope. Compared to open surgery, the field-of-view (FOV) of laparoscopic surgery is narrow, thus the entire surgical workspace cannot be viewed simultaneously. Providing optimal visualization and orientation of the surgical field remains a significant challenge for the human assistant, as verbal communication of the visual intention is not always easy. For example, failure of the assistant to maintain the non-insulated tool tips in the FOV can lead to unrecognized electrosurgical injuries [1]. These visualization challenges can place a greater mental workload on minimally invasive surgeons [2]. The recognized importance of good camera handling and navigation is clear from its recent inclusion in training curricula for surgical residents [3]. Despite these, overcoming problems with assistant fatigue, tremor and close proximity to the surgeon for certain procedures cannot be avoided with training alone.

To address these problems, robotic assisted camera control systems have been commercially developed to assist the surgeon during an operation. These include the voice controlled Automatic Endoscope Optimal Position (AESOP) system from Computer Motion Inc. [4]; the EndoAssist [5] which is controlled by a head-mounted infrared emitter on the user's head, which later was redesigned to have a smaller form factor (FreeHand system, from Prosurgics); and the finger joystick controlled SoloAssist from AktorMed [6].

The purpose of this paper is to introduce a gaze contingent robotic camera control system based on real-time gaze gestures. The main challenge in using gaze contingent control is the difficulty to design an intuitive control interface that enables the user to convey their intention without being affected by aberrant or idiosyncratic saccadic eye movements. The eyes are traditionally used as an information gathering function [7] rather than an input source, for example to control a robotic laparoscopic arm, and it is a challenge to distinguish solely from their Point-of-Regard (PoR) whether they want the camera to zoom, pan, or remain stationary. To overcome this problem, one could use an external input source such as a button press or a pedal switch [8]. However in an operating theatre with potentially several pedals, introducing more foot-switches can lead to instrument control clutter. When reverting to gaze only methods, previous research has utilized dwell-time on fixed regions of the screen to convey panning [9]. Unfortunately, such methods can result in the use of large amounts of screen coverage and further screen area would be needed to introduce zoom functionality. Blink detection, which was previously used in gaze typing [10], could be another method to capture user intention during camera control. However requiring the surgeon to close their eyes for a fixed amount of time could be dangerous for the patient.

One area that has not been explored in surgical camera control is the use of gaze gestures. Gaze gestures can use characteristic eye movements to trigger multiple camera control modes, *e.g.*, panning and zooming. To our knowledge, this is the first time gaze gesture recognition has been used to control a robotic arm for minimally invasive surgery. One significant advantage of gaze gestures is their robustness to eye tracker calibration shifts. Gaze gestures have been previously used for gaming [11], eye typing [12] and Human Computer Interaction (HCI) [13]. In this paper, we have used Hidden Markov Models (HMMs) for gaze gesture recognition. Based on this, multiple input commands can be learned to empower the surgeon to maintain full

Kenko Fujii, Antonino Salerno, Kumuthan Sriskandarajah, Ka-Wai Kwok, Kunal Shetty, and Guang-Zhong Yang are with the Hamlyn Centre for Robotic Surgery, Imperial College London, SW7 2AZ, London, UK (email: k.fujii09@imperial.ac.uk).

control of the camera, whilst simultaneously enabling both hands to focus on the operation. With laparoscopy, the surgeon's visual control and feedback is removed. In the proposed system, the important role of the surgeon's eyes, (*i.e.*, localizing, tracking and orienting) are restored back into the 'control loop' of the camera.

In the following section, the HMM based gaze gesture training and recognition algorithm will be explained (subsections A-C), followed up by the robot control (subsection D). In section III, the experimental setup, the system user interface (UI) and the experimental protocol will be delineated. Section IV discusses the results and the paper finishes with the conclusion and further work discussed in Section V.

II. METHODS

The overall setup of the gaze contingent camera control system is illustrated in Fig. 1. Standard subject-specific eye calibration is performed so that accurate PoR is attainable. Afterwards, the remote center of motion (RCM) for the laparoscope is set, which is then inserted into the phantom, and the surgeon can promptly start using the robotic laparoscopic system afterwards. In order to move the camera, one of the two gaze gestures is performed: "pan" or "zoom". The use of gaze gestures eliminates the need to have an external input mechanism such as a button or foot pedal, whilst enabling the operator to perform a bimanual task simultaneously without the need for a camera assistant.

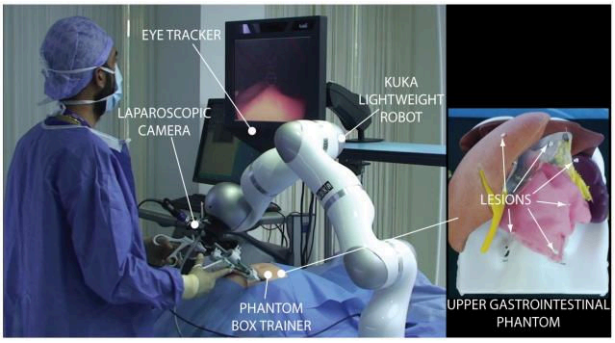

Fig. 1: A surgeon using the gaze contingent camera system. The operator is able to simultaneously navigate and perform a bimanual surgical task.

A. Gaze Gestures: Using Hidden Markov Models

HMMs represent stochastic sequences where the states, rather than being directly observed, are associated with a probability [14]. An HMM is typically represented by a set of N states $S = \{s_1,...,s_N\}$, which are interconnected to each other, and a number of k discrete symbols $V = \{v_1,...,v_k\}$. It is described by an observation sequence $O = \{O_1,...,O_M\}$, a transition matrix $E = \{e_{ij}\}$, which represents the transition probability from state i to state j as $e_{ij} = P(q_{t+1} = S_j \mid q_t = S_i)$, where $1 \le i, j \le N$ and q_t is the state at time t, and the emission probability matrix $F = \{f_{jk}\}$, where f_{jk} is the probability of generating the symbol v_k from the state q_j, with $f_{jk} = P(v_k at\ t \mid q_t = S_j)$, where $1 \le j \le N$ and $1 \le k \le M$. The initial state probability distribution is denoted by $P_i = \{p_j\}$, where $j = 1,2,...,N$. The learning part of HMMs consists of defining an initial state probability P and specifying the optimal state transition and emission probabilities, given a set of observations O. The parameters that most probably describe the set of observation are iteratively defined by using the Viterbi algorithm. Given the trained model and an observation sequence, the probability that the given observation sequence is described by the model is calculated by the Forward-backward algorithm.

A single HMM represents a single gesture and as many HMMs can be added as necessary depending on the number of gestures that need to be incorporated into a given system. For the system design presented here, two gestures need to be utilized, but extending the system to add more gestures would be straightforward. The two gestures here are modelled using 'left-to-right' HMMs; a type of HMM that is often applied to dynamic gesture recognition as the state index transits only form the left to right as time increases [14]. The "zoom" gaze gesture is defined by the following sequence of eye movements; gaze at the center of the screen, then to the bottom left corner, back to the center, and finally back to the bottom left corner. This effectively is a three-stroke gesture [11]. Conversely, the "pan" gaze gesture is the similar to the "zoom" gaze gesture but instead, the user is required to look at the bottom right corner of the screen. Gaze gestures towards the corner of the screen were chosen to minimize obstruction of the camera view, minimize occupying screen real-estate and reduce detection of false gaze gestures. The two gesture's trajectories are illustrated in Fig. 2.

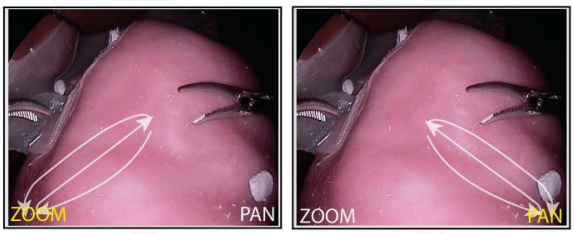

Fig. 2: Example illustration of the (a)"zoom" and (b)"pan" gaze gestures used for this study.

B. Gaze Gestures: HMM Training

One of the challenges in a gaze gesture recognition system is to enable the system to recognize the different intended gestures accurately (*i.e.* high recognition accuracy) and discriminate against natural visual search activity (*i.e.* low false positive rate). Hence, both intentional and unintentional gaze gesture data were used to define the model parameters. 300 intentional gaze gesture sequences for the two gaze gestures were collected from ten engineering students. From the same ten subjects, unintentional gaze gesture data was collected during a five minute web browsing task.

In order to train the HMMs, a discrete codebook that captures the relevant features for the gaze gestures of interest needs to be specified. To achieve this, gaze gesture trajectory data are clustered using a k-means algorithm. The optimal number of clusters for the two gestures is set to 5. The codebook consists of the symbol number, centroid coordinates of each cluster and the radius of each cluster. Symbol numbers are assigned by using the distance between

the observation and the centroid of each cluster provided that it is within the defined radius. If the observation is outside the feature space, it is discarded.

A 10-fold cross validation was run across HMMs of different number of states (from 2 up to 8 states, a total of 6 runs) using the 300 gaze gesture training data sequences. During each training instance a detection probability threshold was defined to be at a 95% confidence limit of the training data sequences' inference values outputted from the trained HMM. The overall recognition accuracy of the system was then defined by using this threshold on the testing data.

Fig. 3(a) illustrates the recognition accuracy and false positive rate of the system across the increasing number of states used in the HMM for the "pan" gaze gesture and Fig. 3(b) for the "zoom" gaze gesture. It was decided that 6 states were to be used for the HMMs as this showed a good trade-off between accuracy, robustness and complexity of the system.

All 300 training data sequences were used during training of the gaze gesture HMMs. Fig. 3(c) and Fig. 3(d) show the inference value histograms obtained from the gaze gesture testing data using 6 state HMMs along with unintended gaze gesture data (natural browsing data and the alternative gesture data) for the "pan" and "zoom" gaze gesture respectively.

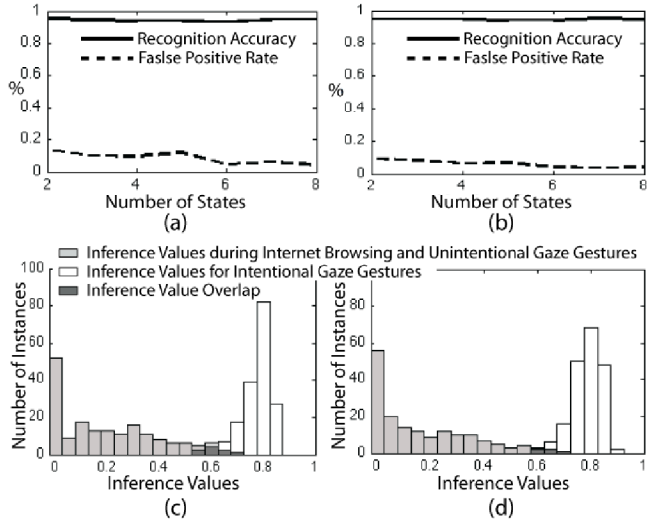

Fig. 3: The recognition accuracy and false positive rate across different number of states for the "pan" gaze gesture (a) and "zoom" gaze gesture (b). The inference values of unintended gaze gestures and the respective inference values of intended gaze gestures for (c) "pan" gestures and (d) "zoom" gestures respectively.

The white bars in the histogram show the inference values of the unintended gaze gesture data, whilst the dark grey values identifies the overlap of inference values between the intended and unintended gaze gesture data. From Fig. 3(c-d) a cutoff threshold for the inference values is observable, making it possible to differentiate the intentional and unintentional gaze gestures. An inference threshold of 0.65 was used for the "pan" gaze gesture detection and an inference threshold of 0.7 was used for the "zoom" gaze gesture detection.

C. Gaze Gestures: On-line Recognition Algorithm

The algorithm shown in Fig. 4 is proposed to enable real-time gaze gesture recognition. A de-noise median filter with a 150ms time window is applied to the PoR outputs from the eye tracker, from which gaze gestures are then segmented. Since the nature of the three-stroke gaze gestures used for our system requires the user to produce three sequential fast eye movements, *i.e.* saccades. The gesture segmentation algorithm was designed to detect three saccades. A typical saccade can be detected by searching for two sequential gaze data samples that exceed a velocity threshold of $300^{\circ}s^{-1}$. A velocity threshold saccade detection method based on Salvucci *et al.* [15] was implemented whilst temporal constraints were introduced so that the gaze gestures were not excessively long (each gaze stroke was required to be less than 750ms). This was placed to ensure that the segmented gaze gesture were both continuous and intended by the user. Furthermore, it was identified that any segmented gaze gesture could potentially contain part of an intended gaze gesture (*i.e.* one third or two thirds of the intended three-stroke gaze gesture). Therefore, instead of discarding a segmented three-stroke gaze gesture sequence after it failed to be recognized as a gesture, buffers were used to store the last two strokes of any segmented potential gaze gesture. The next incoming gaze stroke would be then added onto the previous two strokes in the gaze sequence. After a gaze gesture has been recognized, this buffer would be cleared and the algorithm would search for a new three-stroke gaze gesture sequence again.

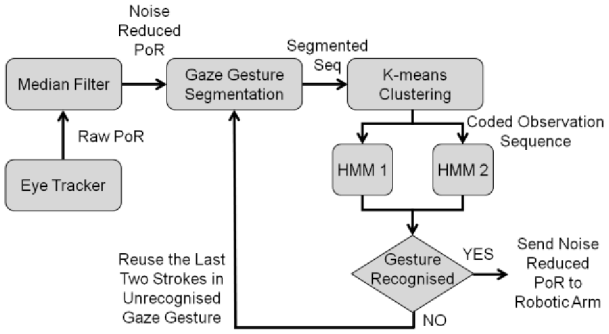

Fig. 4: The online gaze gesture recognition algorithm. The PoR's are median filtered, gaze gestures segmented and coded. The sequence is tested for inference of both gaze gesture HMMs. If a gaze gesture is recognised then PoR information is sent to the robotic arm.

Once a gaze gesture sequence is segmented, it is converted into an observation sequence using the k-means cluster codebook determined previously from training. This new coded sequence is then tested for the respective "pan" (HMM1) and "zoom" (HMM2) gestures. The recognized gesture is the one with the maximum inference value from HMM1 and HMM2, given that it is above the inference value threshold defined by the training. If either one of the gaze gestures is recognized, the PoR is sent to the robotic arm in order to control it as described further in Section III *B*, otherwise no input is given to the robotic laparoscope. The next section will explain the control scheme of this robotic arm in further details.

D. Robot Control

The general control scheme of the robotic camera manipulator has been based on a Cartesian impedance

controller that computes the command torque τ_c for each joint according to the position error e in Cartesian space and the compensation of the whole robot dynamics. The aforementioned error has been estimated as the difference between a reference pose x_d provided by a minimum jerk trajectory planner and actual pose x retrieved from the robot's forward kinematics. As shown in Fig. 5, the trajectory planner is updated on-line in compliance with the gaze co-ordinates $PoR = [x_{eye} \quad y_{eye}]^T$ to place the camera where the end user is looking.

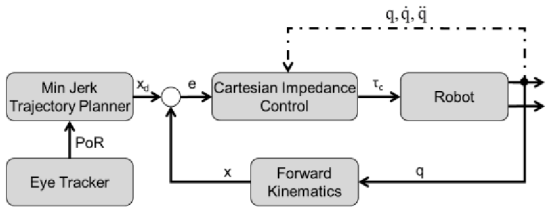

Fig. 5: Block scheme of the control architecture.

The Cartesian impedance control was chosen to guarantee both a safe human robot interaction and intuitive camera positioning during surgery. The command torque vector $\tau_c \in \Re^{n\times 1}$ has been computed according to the following law:

$$\tau_c = J^T K\left(x_d - x\right) + D\left(d_c\right) + F_{dynamics}\left(q,\dot{q},\ddot{q}\right) \qquad (1)$$

where, $J \in \Re^{6\times n}$ is the Jacobian matrix for a robot with n DoFs and n joints, $K \in \Re^{6\times 6}$ is a diagonal stiffness matrix, $x_d \in \Re^{6\times 1}$ is the desired pose, $x = [p, \varphi]^T \in \Re^{6\times 1}$ is the actual pose in terms of position $p \in \Re^{3\times 1}$ and orientation $\varphi \in \Re^{3\times 1}$, $D \in \Re^{n\times 1}$ is a damping vector depending on the damping coefficients d_c, and $F_{dynamics}\left(q,\dot{q},\ddot{q}\right) \in \Re^{n\times 1}$ is the robot dynamics compensation including the terms of inertia matrix, centrifugal and Coriolis torques, friction and gravitational torque vector. $\left(q,\dot{q},\ddot{q}\right) \in \Re^{n\times 1}$ are vectors of joint position, velocity and acceleration respectively. The main benefit in employing this control approach is that the robot's visco-elastic properties during interaction with patient's tissue can be regulated by tuning robot stiffness K and damping $D(d_c)$ matrices according to the robot behaviour like a passive mass-spring-damper system. In the 'camera assistant' mode mentioned in Section III *B*, the control law is given by $\tau_c = F_{dynamics}\left(q,\dot{q},\ddot{q}\right)$. Further details about the Cartesian Impedance control can be found in [16, 17].

Referring to Fig. 6(a), the pose vector x of the camera frame $\{C\}$ with respect to the base frame $\{B\}$ can be reconstructed by the homogeneous transformation matrix:

$$T_c^b\left(q\right) = \begin{bmatrix} R_c^b & p_c^b \\ 0^T & 1 \end{bmatrix} = T_e^b\left(q\right) T_c^e \qquad (2)$$

where $R_c^b \in SO(3)$ and p_c^b are respectively the rotation matrix and origin position of tool frame $\{C\}$ with respect to the base frame $\{B\}$, T_e^b is the homogeneous transformation matrix of the robot end-effector frame $\{E\}$ with respect to the base frame $\{B\}$ based on robot forward kinematics and T_c^e is the fixed homogeneous transformation between the camera frame $\{C\}$ and the end-effector frame $\{E\}$. The rotation matrix R_c^b can been identified by using the angle/axis technique [18].

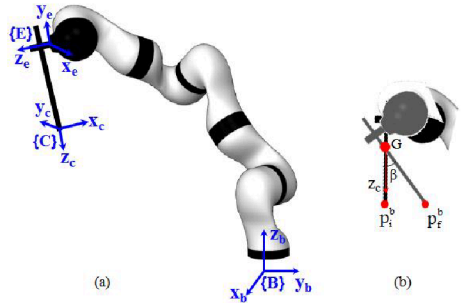

Fig. 6: (a) Reference frames placement; (b) RCM constraint

In order to translate the surgeon's gaze to the robot motion, a minimum jerk trajectory planner has been implemented for the frame $\{C\}$ motion with respect to the base frame $\{B\}$. More specifically, each movement was broken down into a series of finer point-to-point motions to achieve a smoother and safer camera movement in the Cartesian space. The following equation is a parametric representation for a segment connecting (in base frame $\{B\}$) the initial point p_i^b to the end point p_f^b in a time interval t_f.

$$p^b(s) = p_i^b + \frac{s(t)}{\left\| p_f^b - p_i^b \right\|}\left(p_f^b - p_i^b\right) \qquad (3)$$

where the arc length $s(t)$ is the fifth order polynomial function of time:

$$s(t) = \sum_{i=0}^{5} c_i t^i \qquad (4)$$

The point p_i^b represents the actual p_c^b value, whereas the final point p_f^b has been computed by relying on the gaze coordinates. The coordinates $PoR = [x_{eye} \quad y_{eye}]^T$ provide information about the direction of the gaze vector in the eye tracker screen plane as:

$$\alpha = \tan^{-1}\left(\frac{y_{eye}}{x_{eye}}\right) \qquad (5)$$

Considering $\left\| p_f - p_i \right\| = L_d$ as constant, the point p_f^c with respect the reference frame $\{C\}$ can be expressed as:

$$p_f^c = L_d \cdot \begin{bmatrix} \cos(\alpha) & \sin(\alpha) & 0 \end{bmatrix}^T \qquad (6)$$

and the corresponding point p_f^b in base frame $\{B\}$ is given as:

$$\begin{bmatrix} p_f^b & 1 \end{bmatrix}^T = T_c^b \begin{bmatrix} p_f^c & 1 \end{bmatrix}^T \quad (7)$$

Equations (3)-(7) provide a smooth point-to-point motion [19, 20] in the x-y plane of the reference frame $\{\mathrm{C}\}$ in Fig. 6 (a) with respect to base frame $\{\mathrm{B}\}$. With reference to Fig. 6 (b), during a minimally invasive laparoscopic procedure, the access trocar provides a fixed insertion point that acts as a RCM. Therefore the motion planning of the camera tool orientation can be constrained by evaluating the following rotation matrix by means of angle/axis technique [18]:

$$R(\beta,\hat{a}) = \begin{bmatrix} \hat{a}_x^2(1-c_\beta)+c_\beta & \hat{a}_x\hat{a}_y(1-c_\beta)-\hat{a}_z s_\beta & \hat{a}_x\hat{a}_z(1-c_\beta)+\hat{a}_y s_\beta \\ \hat{a}_x\hat{a}_y(1-c_\beta)+\hat{a}_z s_\beta & \hat{a}_y^2(1-c_\beta)+c_\beta & \hat{a}_y\hat{a}_z(1-c_\beta)-\hat{a}_x s_\beta \\ \hat{a}_x\hat{a}_z(1-c_\beta)-\hat{a}_y s_\beta & \hat{a}_y\hat{a}_z(1-c_\beta)+\hat{a}_x s_\beta & \hat{a}_z^2(1-c_\beta)+c_\beta \end{bmatrix} \quad (8)$$

where β is the rotation angle around the axis $\hat{a} = [\hat{a}_x, \hat{a}_y, \hat{a}_z]^T$ in the base frame.

The rotation matrix $R(\beta,\hat{a})$ describes the rotation of the camera tool around RCM constraint in order to plan the camera movement from point p_i^b to point p_f^b. Given the vector $\rho = p_f^b - p_G^b$ in Fig. 6(b), with p_G^b position with respect the base frame $\{\mathrm{B}\}$ of the point G where the RCM is placed, the axis $\hat{a} = [\hat{a}_x, \hat{a}_y, \hat{a}_z]^T$ and the angle β in (8) are

$$\beta_f = a\tan 2\left(\frac{\|z_c \times \rho\|}{\|\rho\|}, \frac{z_c^T \cdot \rho}{\|\rho\|} \right), \ \hat{a} = \frac{z_c \times \rho}{\|z_c \times \rho\|}, \quad (9)$$

where z_c is the z- axis unit vector of frame $\{\mathrm{C}\}$.

Therefore, the homogeneous transformation matrix of the camera frame, with respect to the base frame in the initial and final configuration respectively ${}^iT_c^b$, ${}^fT_c^b$ will be:

$${}^iT_c^b = \begin{bmatrix} {}^iR_c^b & p_i^b \\ 0^T & 1 \end{bmatrix}, \ {}^fT_c^b = \begin{bmatrix} R(\beta,\hat{a})\,{}^iR_c^b & p_f^b \\ 0^T & 1 \end{bmatrix}, \quad (10)$$

where ${}^iR_c^b$ is the rotation matrix of the frame $\{\mathrm{C}\}$ at the instant time $t = 0$. The timing law (4) has been also assigned to β with $\beta(t)\big|_{t=0} = 0$ and $\beta(t)\big|_{t=t_f} = \beta_f$ in order to simultaneously translate (3)-(7) and rotate (8)-(10) the frame $\{\mathrm{C}\}$ in compliance with RCM constraint. The procedure described thus far addresses the motion modality of camera panning in Section III *B*.

With reference to TABLE I, three different camera velocities based on the normalised Euclidean distance r of the gaze from the centre of the screen have been selected for the "pan" mode according to clinical requirements and the aforementioned assumption of small movements.

Without a loss of generality, the same motion generation approach was applied to "zoom" mode by considering

$$p_f^b = p_f^b \pm z_c \cdot L_d \ , \ {}^fR_c^b = {}^iR_c^b \quad (12)$$

where the $\pm$ was used to zoom out and in respectively. The variables L_d and t_f have been fixed to 5.0 mm and 0.500 s respectively for zooming action.

TABLE I. MOTION PARAMETERS FOR EACH ROBOT MOVEMENT

	Planner parameters	Avg. speed	Condition
Region 1	L_d =0.0 mm t_f = 0.010 s	0.00 mm/s	$r < r_1 = 0.18$
Region 2	L_d =5.5 mm t_f = 0.325 s	16.9 mm/s	$r_1 \le r < r_2 = 0.38$
Region 3	L_d =7.0 mm t_f =0.300 s	23.3 mm/s	$r > r_2$

III. IMPLEMENTATION

A. Experimental Setup

The experimental platform which is illustrated in Fig. 1 includes a Tobii 1750 eye tracker, a 10mm zero degree Stortz laparoscope, a Storz Tele Pack light box, a Kuka Light Weight Robot (LWR) [16] and an upper gastrointestinal phantom with simulated white lesions.

Both the robot and the HMM gaze gesture recognition algorithms have been implemented in C++ and run at 200 Hz and 33.3Hz respectively. All experimental data including the HMM gaze gestures, the camera-view video, the PoR and camera position in Cartesian space obtained from the robot forward kinematics, were recorded at 33.3Hz with respective time stamps. The bi-directional communication and synchronization between the devices was achieved using the UDP protocol via an Ethernet cable.

B. Gaze Contingent Control Modes and UI

During the experiment, three control modes were to be used, namely; 1) Gaze gesture control; 2) Pedal activated control, where the user activates the camera control with a foot pedal but directs the camera with their PoR; and 3) Camera assistant mode, the assistant controls the camera for the participant. The experimental participant would need to communicate to the assistant how they would like the camera to be moved.

The gaze gesture control UI shown in Fig. 7 is designed to enable the user to intuitively switch between the "pan", "zoom" and stopping of the camera. At the outset, the camera system is stationary and is waiting for a gesture input from the user. The user will be observing a camera view as in Fig. 7(a) where there are guidance "zoom" and "pan" text in the left and right bottom corner respectively. Whilst using the system, the user is able to see their PoR on the screen in the form of a white dot. They also have the option to turn this off. If a "zoom" gaze gesture is triggered then the UI switches from Fig. 7(a) to Fig. 7(b) and the user will be able to zoom in by looking anywhere above the horizon and zoom out by looking below. Guidance text is also overlaid on the camera view and the camera can be stopped by looking into the white circle in the centre for 750ms. In contrast, if a "pan" gaze gesture is inputted by the user, the UI switches from Fig. 7(a) to Fig. 7(c) where the camera pans in the vector direction of the gaze from the screen center.

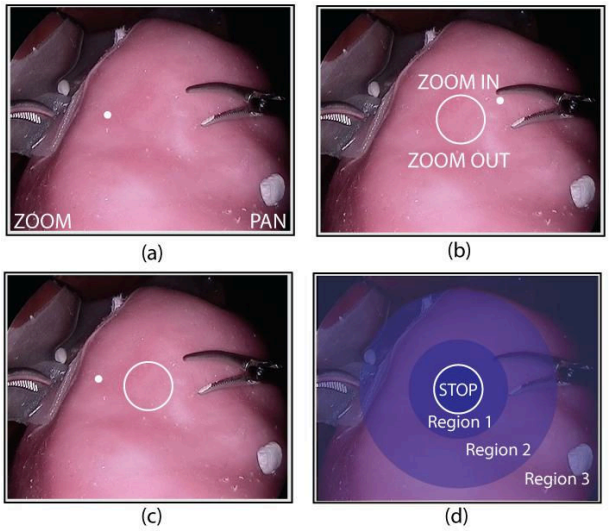

Fig. 7: Gaze gesture controlled robotic laparoscopic UI. (a) The UI waits for the user to perform a "zoom" or "pan" gaze gesture. If a "zoom" gaze gesture is detected then the user will be taken to (b), or if a "pan" gesture is detected it will take them to (c). While in (b) or (c) the camera can be stopped by looking into the white circle for 750ms. After the camera stops, the user is taken back to (a). (d) Illustration of the camera velocity and respective regions under "pan" camera control mode.

If the user's PoR is within a region of $r \leq 0.1$ from the screen center, then the camera will stop after 750ms. In "pan" mode, the screen area is separated in three radial regions where gazing within each region would prompt the camera to move at a different velocity. This region separation is illustrated in Fig. 7(d) with respective radial region values described as in TABLE I. If the PoR is within region 1, the camera will remain static. This region was introduced to enable the surgeon to have a stable working area to operate whilst maintaining an active "pan" mode. If the PoR falls within region 2 and region 3, the camera pans in a medium and fast velocity set at 16.9 mm/s and 23.3 mm/s respectively. The different speed regions were introduced to enable an intuitive control of the camera whilst maintaining a known maximum velocity that would be safe for the patient.

In order to address the potential safety issue when the eye tracker looses tracking of the user's eyes, an effective safety mechanism was introduced where the robotic laparoscopic holder would immediately stop under lost gaze tracking conditions. If the user's gaze is re-detected, then the robotic laparoscope would resume in the same user control mode.

The pedal activated control mode is similar to the UI of the gaze gesture control except that the onset camera screen view shown in Fig. 7(a) would not have the "zoom" and "pan" text in the corner as a foot pedal was being used to activate the camera arm. During pedal activation control mode, a 2-lever foot pedal is used. To activate the "zoom" or "pan" mode the user is required to step on the left or right pedal respectively. Navigation of the camera is conducted using gaze with the same UI as in Fig. 7(b) and Fig. 7(c) for "zoom" and "pan" mode respectively. During camera assistant mode, an operator manually controls the laparoscopic camera using the method explained in Section II D.

C. Experimental Protocol

Eleven surgical residents with a postgraduate year between 3-7 (PGY3-7, male=10, female=1) were consented to participate in this study. The laparoscopic experience was 536 (+/- 315) cases). All subjects were initially trained on use of both the gaze gesture and pedal activated systems on an abstract navigation task. This was done to mitigate the potential confounding effects of learning when performing the subsequent study task.

The abstract training task involved navigating the robotic laparoscope within a conventional box trainer to find numbers on a 4x5 grid in ascending magnitude. The numerals were of differing font sizes, necessitating zooming and panning to successfully complete the task. Training was completed when they had met a minimum baseline proficiency task completion time, when there was no further improvement in completion time and they could reproduce their best time on three consecutive occasions.

A modified upper gastrointestinal staging laparoscopy phantom was used in an immersive laparoscopic box trainer. The subjects were tasked with identifying and "biopsying" (*i.e.* removing) a set number of randomly placed lesions on the phantom. Importantly, this simulated task required subjects to use a bimanual technique, with one instrument either manipulating or retracting tissue, whilst the other removes the lesion.

The sequence in which subjects performed the various task modes was randomized to mitigate the learning effects on the phantom. The three modes were: i) conventional human assistant, ii) gaze gesture, and iii) foot pedal activation. The same human camera assistant was used throughout the study for all subjects. The assistant was experienced in control of the robotic laparoscope arm.

Each task was assessed quantitatively, with task time measured in seconds and camera path length in meters. Each subject performed the task with each modality twice. Following each task they completed the National Aeronautical Space Agency–Task Load Index (NASA-TLX) questionnaire. This is a well validated subjective questionnaire comprising of six variably weighted parameters that contribute to task workload [21].

At the end of the trial, subjects completed a questionnaire providing their laparoscopic experience and rated how difficult it was to learn to use a) the gaze gesture system and b) the foot pedal activation system on a visual analogue scale.

IV. Results and Analysis

In order to assess the performance of the new gaze contingent laparoscopic camera control system we looked at the following combination of quantitative and qualitative system and usability performance measures;

- Recognition accuracy of the HMM gaze gestures
- Camera path length – to assess efficiency and ability to use the system.
- Task completion time – to assess usability.
- NASA TLX questionnaire – validated measure of subjective workload for each of the three control modes.

- Visual Analogue Score – rating difficulty of system skill acquisition.

Statistical analysis was performed using IBM© SPSS© v19. Normality tests were initially performed, followed by Mann Whitney U test for nonparametric continuous variables between modalities. A 'p value' <0.05 was considered significant. Results are represented as medians with interquartile ranges in parentheses, unless otherwise stated.

A. *Accuracy of Online HMM implementation*

Occurrence of false positive (*i.e.* when the user does not perform a gaze gesture, but the algorithm triggers a gaze gesture) and false negative (*i.e.* when the user performs a gaze gesture, but it is not recognized by the algorithm) gaze gestures during the experiment were counted via *post-hoc* observation of the recorded camera-view videos. The recognition accuracy is calculated by dividing the number of true positive gaze gestures by the sum of the true positive and false negative gaze gestures. The false positive rate is obtained by dividing the number of false positive gaze gestures by the sum of the number of true negative and false positive gaze gestures during the gaze gesture control mode of the experimental task. The results are summarized in TABLE II.

TABLE II. RECOGNITION ACCURACY AND THE FALSE POSITIVE RATE OF THE HMMS DURING THE EXPERIMENTAL TRIAL

	Recognition Accuracy	***False Positive Rate***
HMM1 ("Pan")	95.6%	2.2 %
HMM2 ("Zoom")	98.3%	0.6 %

Low recognition accuracy would cause frustration for the surgeon, as his/her gaze gestures are not being accepted, resulting in an experiential delay in being able to move the camera. A high false positive rate would lead to the camera starting to move without the surgeon intending, resulting in the need to correct the camera position. These factors, if significant, would lead to a poor uptake of our system. The average recognition accuracy for the HMMs was 97.0% whilst the average false positive rate was 1.4%. This demonstrates that the use of HMMs for gaze gestures has the potential to be both user-friendly and safe to be used during camera control of the robotic laparoscopic arm.

B. *Quantitative Analysis - User performance tables*

For this study, each of the eleven subjects met the baseline proficiency and training requirements. Two of the subjects wore glasses and three wore contact lenses.

The camera path length was measured in order to assess how efficient and user-friendly our system is compared to the other two modes. As illustrated in Fig. 8(a), a significantly shorter camera path length for the gaze gesture modality compared to the assistant was observed with median and interquartile range values of (0.896m [0.87] vs. 1.710m [1.26]; p=0.037). The pedal activation control mode also showed statistically significant shorter camera path length when comparing it to the camera assistant (1.076m [0.88] vs. 1.710m [1.26]; p=0.031). Fig. 9 illustrates an example camera path of a single subject during the gaze gesture based control of the camera (left) and the camera assistant on the (right). It is clear that the camera assistant shows a more volatile camera trajectory. In contrast, the camera path during the gaze gesture based gaze contingent control mode is smooth and still offers the same navigational workspace of an assistant. An unexpected observation was the statistically insignificant difference (p=0.690) in comparative task completion times across the three modalities. This indicates that our gaze contingent system is as competent as using a camera assistant, which is the current gold standard.

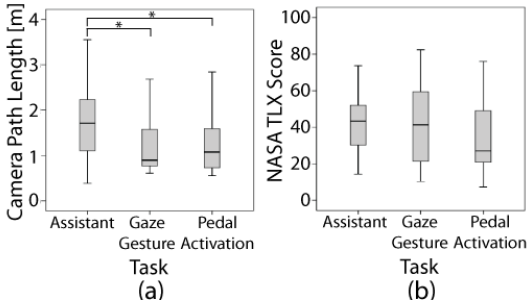

Fig. 8: (a) Camera path length (m), (b) the NASA TLX scores for all control modes. Significant p values shown with a * mark above bar charts.

Our system was also assessed for its contribution to the workload of the surgeon with the NASA-TLX questionnaire, in which cognitive workload is an assessed factor. It is desirable that any surgical innovation does not add to the cognitive burden of the surgeon. Importantly, there was no statistically significant difference in NASA-TLX scores for the gaze gesture modality versus the human assistant (p=0.972) or the foot pedal (p=0.217) as shown in Fig. 8(b).

Visual analogue scores for rating subjective difficulty in skill acquisition between our gaze gesture system and the pedal activation control system were not significantly different (0.853 [0.15] vs. 0.884 [0.14]; p =0.178). 100% of surgeons rated both systems as 'easy' to learn. Pedals are considered commonplace in the operating theatre. Thus, for gaze gestures to be as easy to learn as the foot pedal is a measure of the intrinsic ergonomics of gaze gestures. The user performance statistics are presented in TABLE III.

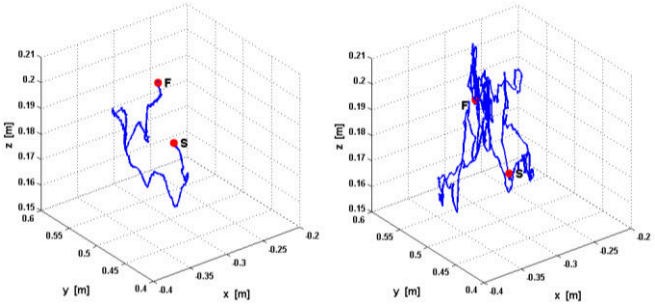

Fig. 9: Camera path for a given subject during gaze gesture based gaze contingent (left) and assistant (right) control of the laparoscope.

In the questionnaire, the subjects were asked to state their preferred mode and 91% of surgeons (10/11) documented they would use the gaze gesture system in clinical practice and also recommend it to their surgical colleagues.

TABLE III. USER PERFORMANCE TABLE. MEDIAN AND INTERQUARTILE RANGE IN PARENTHESES

	Task Time (sec)	*NASA -TLX Score*	*Cam Path Len.(m)*	*Subj. Skill Acq. Difficulty†*
Gaze Gesture	281.0 [172]	41.33 [39.08]	0.896 [0.87]	0.853 [0.15]
Pedal Activation	265.0 [160]	27 [28.32]	1.076 [0.88]	0.884 [0.14]
Assistant	297.3 [181]	43.33 [23.34]	1.710 [1.26]	-
p value Gaze vs Assistant	0.690	0.972	0.037*	-
p value Pedal vs Assistant	0.460	0.133	0.031*	-
p value Gaze vs Pedal	0.842	0.217	0.814	0.178

*p<0.05, †Visual analogue scale score: 1=very easy, 0=very difficult

V. CONCLUSION AND FUTURE WORK

In this paper, we have provided a novel approach to ergonomically control a 6 DoF camera through the surgeons' PoR (*i.e.* 2 DoF) as the control input via the use of real-time HMM gaze gestures to pan and zoom in the Cartesian space. The ability to reproducibly complete a surgical navigation task has been demonstrated with the gaze gesture modality. The efficiency of the system was demonstrated by the significantly shorter camera path length compared to the current gold standard. Thus, surgeons obtained the desired FOV more accurately and required less fine-tuning. Validation of the HMMs showed that HMMs are effective in recognizing gaze gestures with mean experimental recognition accuracy and false positive rate of 97.0% and 1.4% respectively.

From the results derived, the proposed gaze contingent robotic camera system with built-in safety mechanisms has shown to be ergonomic with no statistically significant difference in cognitive burden when compared to using a camera assistant or when compared to using a foot pedal activated camera system. This means that the participants did not feel that the gaze gestures and gaze control as a complex user modality. This was also reflected in the subjective visual analogue scores. Additionally, the system has the desirable feature of enabling the surgeon to achieve comparable performance without the need of additional foot pedals.

In future work, the surgical task videos will be rated for 'quality' of surgery by independent blinded experts using a validated rating system. This will provide information regarding whether the system has an impact on tool accuracy, efficiency and the number of surgical errors. We also plan to modify the "zoom" mode to enable simultaneous zooming and panning. To this end, more gaze gestures would have to be added to separate zooming in and out. In addition, through adding more gaze gestures we can create an immersive environment for the surgeon to switch on and off a number of surgical applications intra-operatively, for example, augmented reality visualizations to help localize tumors. Further work to recognize spatially invariant gaze gestures is another area under research as well as to implement safety-boundaries for the robot workspace inside patient's abdomen based on pre-operative model of the anatomical site.

REFERENCES

[1] C. C. Nduka, *et al.*, "Cause and prevention of electrosurgical injuries in laparoscopy," *Journal of the American College of Surgeons,* vol. 179, pp. 161-170, 1994.

[2] B. Zheng, *et al.*, "Quantifying mental workloads of surgeons performing natural orifice transluminal endoscopic surgery (NOTES) procedures," *Surgical endoscopy,* vol. 26, pp. 1352-1358, 2012/05/01 2012.

[3] S. Shetty, *et al.*, "Construct and face validity of a virtual reality–based camera navigation curriculum," *Journal of Surgical Research,* vol. 177, pp. 191-195, 2012.

[4] B. M. Kraft, *et al.*, "The AESOP robot system in laparoscopic surgery: Increased risk or advantage for surgeon and patient?," *Surgical Endoscopy And Other Interventional Techniques,* vol. 18, pp. 1216-1223, 2004/08/01 2004.

[5] S. Kommu, *et al.*, "Initial experience with the EndoAssist camera-holding robot in laparoscopic urological surgery," *Journal of Robotic Surgery,* vol. 1, pp. 133-137, 2007/07/01 2007.

[6] K. Sriskandarajah, *et al.*, "Robotic Tele-Manipulating Devices for Laparoscopy Improve Surgical Performance in Simulated Porcine Laparoscopic Cholecystectomies on the ELITE Simulator," *Proceedings of the 2012 Hamlyn Symposium on Medical Robotics,* pp. 36-37, 2012.

[7] G. Z. Yang, *et al.*, "Visual search: psychophysical models and practical applications," *Image and Vision Computing,* vol. 20, p. 291, 2002.

[8] C. Staub, *et al.*, "Human-computer interfaces for interaction with surgical tools in robotic surgery," in *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*, 2012, pp. 81-86.

[9] D. P. Noonan, *et al.*, "Gaze Contingent Control for an Articulated Mechatronic Laparoscope," in *IEEE International Conference on Biomedical Robotics and Biomechatronics*, 2010.

[10] M. Fejtová, *et al.*, "Controlling a PC by Eye Movements: The MEMREC Project," in *Computers Helping People with Special Needs*. vol. 3118, K. Miesenberger, *et al.*, Eds., ed: Springer Berlin Heidelberg, 2004, pp. 770-773.

[11] H. Istance, *et al.*, "Designing gaze gestures for gaming: an investigation of performance," presented at the Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, Austin, Texas, 2010.

[12] J. O. Wobbrock, *et al.*, "Longitudinal evaluation of discrete consecutive gaze gestures for text entry," presented at the Proceedings of the 2008 symposium on Eye-Tracking Research & Applications, Savannah, Georgia, 2008.

[13] D. Rozado, *et al.*, "Low cost remote gaze gesture recognition in real time," *Applied Soft Computing,* vol. 12, pp. 2072-2084, 2012.

[14] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the Ieee,* vol. 77, pp. 257-286, 1989.

[15] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols," presented at the Eye tracking research & applications, 2000.

[16] A. Albu-Schäffer, *et al.*, "The DLR lightweight robot: design and control concepts for robots in human environments," *Industrial robot,* vol. 34, p. 376, 2007.

[17] A. Albu-Schäffer, *et al.*, "A Unified Passivity-based Control Framework for Position, Torque and Impedance Control of Flexible Joint Robots," *The international journal of robotics research,* vol. 26, p. 23, 2007.

[18] B. Siciliano, *et al.*, "Robotics: modelling, planning and control," *Springer,* 2009.

[19] K. J. Kyriakopoulos and G. N. Saridis, "Minimum jerk path generation," in *IEEE International Conference on Robotics and Automation, 1988. Proceedings., 1988*, 1988, pp. 364-369 vol.1.

[20] L. Zollo, *et al.*, "Submovement composition for motion and interaction control of a robot manipulator," in *Biomedical Robotics and Biomechatronics (BioRob), 2010 3rd IEEE RAS and EMBS International Conference on*, 2010, pp. 46-51.

[21] B. Zheng, *et al.*, "Workload assessment of surgeons: correlation between NASA TLX and blinks," *Surgical endoscopy,* vol. 26, pp. 2746-2750, 2012/10/01 2012.