# GRACE: online Gesture Recognition for Autonomous Camera-motion Enhancement in robot-assisted surgery

Nicolò Pasini[1], Andrea Mariani[2], Anton Deguet[3], Peter Kazanzides[3] and Elena De Momi[1]

***Abstract*— Camera navigation in minimally invasive surgery changed significantly since the introduction of robotic assistance. Robotic surgeons are subjected to a cognitive workload increase due to the asynchronous control over tools and camera, which also leads to interruptions in the workflow. Camera motion automation has been addressed as a possible solution, but still lacks situation awareness. We propose an online surgical Gesture Recognition for Autonomous Camera-motion Enhancement (GRACE) system to introduce situation awareness in autonomous camera navigation. A recurrent neural network is used in combination with a tool tracking system to offer gesture-specific camera motion during a robotic-assisted suturing task. GRACE was integrated with a research version of the da Vinci surgical system and a user study (involving 10 participants) was performed to evaluate the benefits introduced by situation awareness in camera motion, both with respect to a state of the art autonomous system (S) and current clinical approach (P). Results show GRACE improving completion time by a median reduction of** $18.9s$ **(**$8.1\%$**) with respect to S and** $65.1s$ **(**$21.1\%$**) with respect to P. Also, workload reduction was confirmed by statistical difference in the NASA Task Load Index with respect to S (**$p < 0.05$**). Reduction of motion sickness, a common issue related to continuous camera motion of autonomous systems, was assessed by a post-experiment survey (**$p < 0.01$**).**

## I. INTRODUCTION

### A. The evolution of vision in surgery

Clinical practice has found in the last two decades a trustworthy ally in surgical robots, especially in minimally invasive surgery (MIS), where surgeons particularly benefit from enhancement of instrumentation, along with improved visual perception, dexterity and ergonomics compared to conventional laparoscopic surgery [1]. Many different categories of medical robots are available on the market, such as the well established da Vinci surgical system, dVSS, (Intuitive Surgical, Sunnyvale, CA, USA) which belongs to the surgeon extender category [2].

Surgical techniques and control modalities have changed due to the deployment of robotic assistance. Visualization modalities have been particularly affected by this aspect: one of the main differences is the loss of direct access to the surgical environment and control over tools and camera. During robot-assisted MIS (RAMIS), access to the patient's body is gained through small incisions for camera and tools, similar to laparoscopic surgery. In the latter, an assistant is often needed to facilitate an eased workflow, alleviating the surgeon from the burden of camera motion. With the introduction of a surgical robot in between the surgeon and the patient, the presence of an assistant to manipulate the camera is no longer required. In this situation, the surgeon is expected to control both the camera and operating tools.

Specific consoles have been designed to provide the user with direct control over multiple robotic arms. To allow teleoperation of both instruments and camera, devices such as the dVSS were specifically designed with a tailored foot pedal tray. This allows to quickly switch between tools and camera control, by handling a couple of manipulators. A drawback resulting from these interfaces is the impossibility for a single surgeon to control both camera and tools simultaneously. Due to the high workload, surgeons may settle for a sub-optimal field of view (FOV) [3]. Surgeons might also allow the tools to be temporarily out of view, due to the effort required to reposition the camera, which can lead to injuries to soft tissues or error in the execution of a surgical procedure [3].

### B. Automation in camera navigation

To mitigate the above mentioned disadvantages and relying on robotics control, the autonomous navigation of the camera has been explored in the past. The proposed approaches to autonomous camera navigation can be clustered into *reactive*, *proactive* or *combined* control strategies [3]. A reactive control architecture is defined as a system in which data streams such as eye gaze [4], [5] or tools tracking [6], [7] are used to move the camera in direct response to changes in these inputs. With a proactive approach, the system incorporates knowledge of the surgery and thanks to prediction-based techniques proposes specific camera viewpoints [8], [9]. With the combination of reactive and proactive systems, a combined camera control modality is obtained [10].

However, the proposed works either require semantically rich instructions from the surgeon or do not rely on procedural knowledge to constantly be aware of the procedure in progress and adapt to camera navigation requirements. This latter condition also results in a continuous - often undesired - motion of the camera, which may lead to discomfort and nausea for the operator [6], better known as *motion sickness*.

### C. Situation awareness in surgery

Extracting procedural knowledge from surgical tasks' objective data is a robust research trend. Also, a robotic system can serve as an enabling factor to achieve this goal, since it is

This research was partially supported by the AccelNet project, grant agreement number NSF OISE 1927354.

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, 20133, Italy, (nicolo1.pasini/elena.demomi@mail.polimi.it)

[2]The BioRobotics Institute and Department of Excellence in Robotics & AI, Scuola Superiore Sant'Anna, Pisa, 56121, Italy, (andrea.mariani@santannapisa.it)

[3]Department of Computer Science, Johns Hopkins University, Baltimore, 21218, Maryland, (anton.deguet/pkaz@jhu.edu)

intrinsically capable to collect meaningful data. Such systems in fact capture diverse data streams (e.g., kinematics, video or data from integrated sensors) simultaneously. These data can be used to design artificial intelligent solutions to assist clinicians over multiple tasks, as camera navigation. As suggested in [11], contextual assistance is crucial and should be guided by the user's intent prediction. A robot should learn certain strategies based on examples or even measurements of the movements of a human operator. This kind of notion regarding procedural knowledge can be defined as *situation awareness*.

Situation awareness can intra-operatively support the physicians by reducing the workload while enhancing safety, detecting hazardous surgical events [12], or even performing surgical *gesture recognition*. Indeed, awareness in surgical tasks has been approached by the segmentation of procedures into pre-determined actions. Based on the desired granularity, actions have been divided into *dexemes*, *surgemes*, *activities*, *phases*, *procedures* and *states*, from finest to coarsest [13]. Artificial intelligence (AI) is more and more used to achieve gesture recognition during surgical procedures and it is able to provide targeted feedback regarding the ongoing process [14], [15] as well as automating surgical tasks or sub-tasks [16]. Nevertheless, the combination of both situation awareness and the automation of camera navigation is still an open challenge in surgical robotics.

Gesture recognition methods can be classified based on the data type that is used as input, hence kinematic data, video streams or a combination of both. Hidden Markov Models (HMM) were first employed to classify motion segments using kinematic data [17], [18], lately outperformed by Linear Dynamical Systems (LDS) [19]. The first approach to gesture recognition using both kinematic and video data was proposed with Conditional Random Fields (CRF) [20], but it was only with Deep Neural Networks (DNN) that the research community found a powerful instrument capable of fine-grained surgeme recognition. Temporal Convolutional Networks (TCN) were introduced to capture long-range temporal dependencies [21], thanks to pooling operations, but experienced imprecise identification of surgemes' boundaries. Later on, 3D CNNs [22] or Spatial Temporal Graph Convolutional Networks (ST-GCN) [23] were proposed to efficiently process also higher-dimensional signals, achieving better performance with respect to spatial and spatio-temporal models. Due to their ability to compute predictions sequentially in time, Recurrent Neural Networks (RNN) have been used to capture long-term dynamics in surgical kinematic data. Thanks to their sequential nature, RNNs can handle signals of different lengths in real time, allowing online gesture recognition. To alleviate gradient descent issues and hyperparameter sensitivity, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been widely used, achieving the best classification performance [24]. In addition to that, a multi-task learning approach has been proposed in [25] to both identify surgemes and progress of the surgical task. Furthermore, works based on unsupervised or semi-supervised learning approaches were presented, which tackled large data necessity issues, since labels are required only for testing, obtaining promising results, as described in [13].

Nevertheless, to the best of our knowledge, no solution has been proposed that takes advantage of online gesture recognition to enhance camera navigation with situation awareness.

### D. Research hypothesis

We designed an online Gesture Recognition system for Autonomous Camera-motion Enhancement (GRACE) during RAMIS to introduce situation awareness in autonomous camera navigation. Given the performance shown by RNNs in the literature [24], [25] the system comprises two LSTM models working in parallel to perform online gesture recognition.

We designed and tested GRACE during a suturing task. This type of surgical process segmentation is well-suited to low-level analysis. As described in [26], different suturing sub-tasks require specific camera adjustments, and specific human gaze patterns can be defined when performing suturing tasks [27], depicting salient regions inside the FOV.

The aim of this work is to test the advantages introduced by situation awareness in autonomous camera navigation during a suturing task, by analysing completion time, workload and motion sickness reduction.

We initially implemented a proof-of-concept of the proposed approach in virtual reality [28]. This work preliminarily validated our research hypothesis, showing the benefits introduced by surgical procedure knowledge on camera automation. Thus, it paved the way to design and integrate GRACE on a real robotic surgical system, while also introducing online gesture recognition with recurrent neural networks. Specifically, the da Vinci Research Kit (dVRK) [29] shown in Fig. 1a, an open research platform derived from the first generation dVSS, was deployed to complete the study.

To validate the benefits introduced by GRACE, we performed a user study comparing GRACE both with the current foot pedal camera control modality and a state of the art System for Camera Autonomous Navigation (SCAN) [7].

## II. Methods

In this section, we introduce the GRACE system, first addressing *Situation Aware Camera Motion Automation* and then the *System Validation*. The first subsection describes the segmentation of the suturing task into four surgemes, the LSTM based gesture recognition model and the autonomous camera control architecture. The second subsection presents the experimental setup, the acquisition protocol, the validation of the classifying model and the evaluation metrics.

### A. Situation Aware Camera Motion Automation

*1) Suturing Task:* the proposed work focuses on the suturing task, decomposing it into *surgemes*. Different suturing surgemes require adjustments of the FOV [26] and are associated to specific human gaze patterns [27]. As done by [26], we identify the suturing sub-task as a repetition of: finding

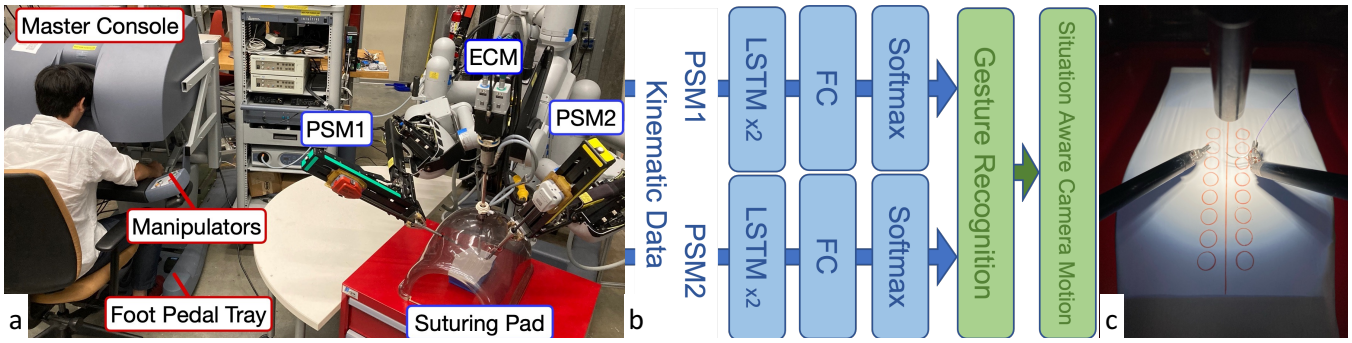


Fig. 1. dVRK setup & GRACE architecture: main components of dVRK system are shown in (a), along with the suturing pad inserted inside a mannequin. The user handles a pair of manipulators to teleoperate two Patient Side Manipulators (PSM) and an Endoscopic Camera Manipulator (ECM). Kinematic data coming from both PSMs is then fed into two parallel recurrent neural networks (b), each composed by two LSTM layers, two Dropout and Batch Normalization layers, followed by a fully connected layer (FC), with a final softmax activation function for gesture classification. Based on the recognized gesture, the camera is moved over the suturing pad, shown in (c).

a suitable position for needle insertion (*Needle Positioning* - NP), push (*Tissue Bite* - TB) and pull (*Suture Throw* - ST) of needle and thread through the tissue. *Reaching for the Needle* - RN - has been added to label cases in which the needle is not handled by the grippers. Note that this refers only to the action of placing the stitching material through the tissue, therefore the knot tying phase is not considered.

*2) Online Gesture Recognition:* to allow camera motion with situation awareness, we devised a deep learning model capable of performing online gesture recognition. This system is based on two LSTM models working in parallel, designed specifically for each arm. They have been trained separately to recognize respectively RN or ST for the left arm, RN, NP or TB for the right arm. The final classification corresponds to the gesture recognized with the highest confidence score from the two models. Given the necessity to work in real-time, we did not include video as input data, due to its higher computational cost and time.

We selected the needle grippers' end effector pose (6), opening angle (1), linear velocity (3), angular velocity (3), joint state (3) and relative distance (1), resulting in a 17 dimensional feature. The time window length is equal to 5 timestamps, resulting in overlapping windows with a data input shape of $5 \times 17$ acquired at 30 Hz.

The model architecture is shown in Fig. 1b: a two-layer LSTM model with respectively 128 and 64 hidden units, with kernel regularizer $l1 = 0.001$, two Dropout layers with $dropoutrate = 0.2$ and 0.3, two Batch Normalization layers with $momentum = 0.99$, followed by a dense layer, with a final softmax activation function for gesture classification. We used the Adam optimizer with a starting learning rate $lr = 0.001$, and at each training iteration we computed the classification loss using the categorical cross-entropy. We selected *early stopping* as the criterion to stop model training, monitoring the loss with a patience of 200 epochs. When learning was stuck for more than 40 epochs, *learning rate reduction* was applied by halving $lr$ until a minimum of $10^{-5}$.

Given the fact that every gesture can be repeated in different positions in space, in order to remove the positional dependency we preprocess the input matrix by subtracting the tool's initial position from each subsequent position. As a result, every classified gesture will have the $(X,Y,Z)$ origin fixed at $(0,0,0)$ for the first timestamp.

*3) Situation Aware Camera Motion:* the autonomous camera motion modality is shown in Fig. 1b and 1c. According to the output of the online gesture recognition, the camera motion system sets the scene center (SC) by relying on the kinematics tracking of the tools' 3D position. Hence, no image segmentation is needed. The camera motion system has been developed on the da Vinci Research Kit, which allows for open access to the robot's kinematics. The laws for camera motion during each gesture were derived from [27], which reports that specific human gaze patterns can be defined when performing a suturing task. Specifically, the camera motion was implemented based on the recognized gesture, as follows:

- Reaching for the Needle: the camera holds a steady position, waiting for the task to start or to hold the needle again after having lost its grip.
- Needle positioning: the camera tracks the weighted SC, as in Fig. 2a. SC is weighted in between the projected tools' midpoint (P) and the patient side manipulator (PSM1) tip, with the aim of minimizing the incidence of motion sickness. The user defines the preferential line for midpoint projection at the task's outset, if necessary, and can modify it during task execution. Pressing a pedal on the foot pedal tray, the user is able to draw a line point-by-point in 3D space using the tip of the surgical tools. Users may proceed without defining a line: in this case P coincides with M, and SC is weighted accordingly.
- Tissue bite: the camera acquires a steady zoomed-in position, as in Fig. 2b, to promote a sharp and fixed FOV over the stitch.
- Suture throw: to conclude the suture, needle and thread must be pulled through the stitch, as in Fig. 2c. For this gesture, the camera follows the same behaviour implemented for *needle positioning* on PSM2.

To allow tools to remain within the FOV, we proposed an adjustable zoom based on the tools' tip distance. Camera's Tip (CT) started at $9cm$ - value set to match the task workspace - from the tracked scene center, and the position

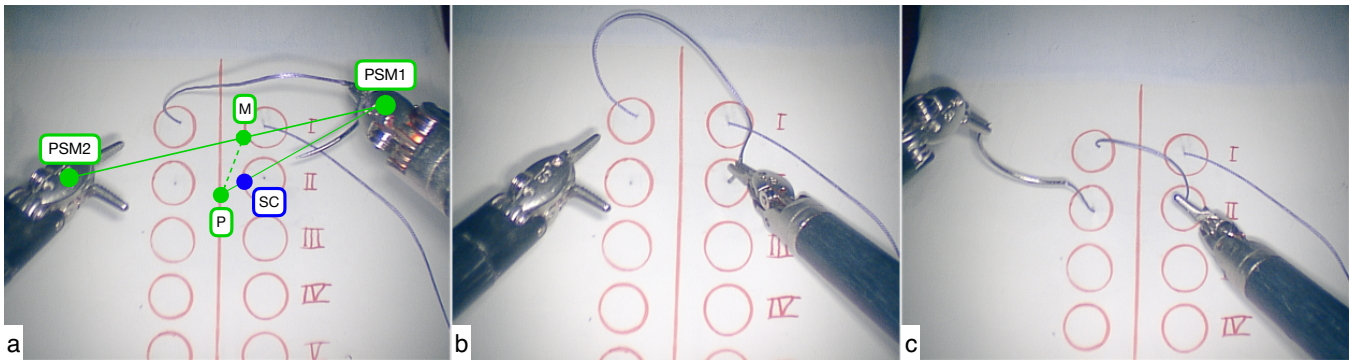


Fig. 2. Gestures & Situation Aware Camera Motion: the suturing environment is shown, with stitches' entry and exit points (red circles) on the suturing pad. (a) Needle Positioning: PSMs tips are tracked and their midpoint (M) is projected (P) on the preferential line, defined by the user at the start of the task. Given the linear shape of the suturing path, the preferential line is straight and ideally aligned with the red path displayed in the figure. Scene Center (SC) is then obtained weighting P and PSM1 position: the higher the weight, the closer SC will be to P. (b) Tissue Bite. (c) Suture Throw: equivalent tracking criteria are used to define SC with respect to PSM2.

was adjusted as follows:

$$CT = SC - (0.09 + z \cdot d)\frac{\mathbf{R}}{|\mathbf{R}|} \tag{1}$$

where SC is the Scene Center in Cartesian space, $\mathbf{R}$ is the vector connecting SC to CT, $z = 0.35$ is a weighting value for zooming range and $d$ represents the distance between the tools' tips. As a result, tools moving apart from each other result in the camera zooming out, allowing for a continuous view of the instrumentation, even when larger FOVs are required. This solution has been introduced to accommodate both narrow views, necessary for TB when tools are in close proximity, and wide views, preferable for both ST and NP.

### B. System Validation

*1) Experimental Setup:* the setup used for our user study is shown in Fig. 1a. The master console of a dVRK is the surgeon's side of the robot. It is equipped with a foot-pedal tray, two Master Tool Manipulators (MTMs) and a stereo viewer for visualization of the surgical environment. On the patient's side of the robot, there are two Patient Side Manipulators (PSMs), holding the surgical tools, and an Endoscopic Camera Manipulator (ECM), holding the camera. A plastic mannequin contains a magnetic pad holding in position a latex non-tear film, as a substitute for soft tissues, with the stitches' entry and exit points as in Fig. 2. The pad can be positioned as desired on the red ferromagnetic surface inside the mannequin, as in Fig. 1a, thanks to two magnets placed under it which allow a steady hold. The last element composing the experimental setup is the surgical needle with thread, used to suture the pad with the help of needle grippers, or Large Needle Drivers, attached to the PSMs.

*2) Acquisition Protocol:* we performed a user study with 10 non-medical users (20 to 27 years, 9 males, 1 female, all right handed). Each user completed the suturing task in 3 modalities: camera navigation with foot pedal control (P), as in clinical dVSS systems, continuous tools' tips midpoint tracking (S), as in [7], and the GRACE system (A).

To complete the suturing task, participants had to pick up the needle and perform a total of 11 stitches, as in Fig. 2. Each camera motion modality was tested 2 times, for a total of 6 repetitions per subject. The FOV proposed at the beginning of every task repetition was such as to not allow a complete view of all the stitches.

The vast majority of previous works relied on the JIGSAWS dataset [17] to train and test their models. However, due to its lack of camera motion, we collected synchronized kinematic and video data during the suturing task performed with the camera foot pedal control modality to build a new dataset, comprehensive of camera adjustments. We later performed offline manual labeling: the resulting dataset, with a dimensionality of more than 150k labelled features windows, was used to train the LSTM based model that we designed to perform situation aware autonomous camera motion (A). The dataset is available at (https://github.com/paso04/Autonomous-Camera-Motion).

As a consequence, all the experiments started with 2 repetitions with the P modality. To allow the process described in the previous paragraph (i.e., labeling and model training), the other 4 repetitions were performed after 2 weeks. This also allowed to minimize any learning effect from the P repetitions. For a direct comparison, the S and A repetitions were performed during the same day. Again, to minimize the influence of any learning effect on the results, permuted block randomization was used to define the order of the 4 repetitions with the S and A modalities for each user.

Before starting the task, each participant was given an introductory speech, in which the main components of the Master Console were described, and 5 minutes of training, during which they could familiarize themselves with the robotic platform. The suturing environment was displayed on the stereo viewer inside the Master Console, in which the users placed their head.

The experiments were carried out after Institutional Review Board (IRB) approval (protocol number: HIRB00000701) with oral consent from participants. The official NASA Task Load Index (TLX) iOS App has been used for measuring the subjective workload.

*3) Online LOUO Validation:* we performed the Leave One User Out (LOUO) approach to validate our model. Validation has been performed online during suturing task completion, for every user. To be able to do so, we trained

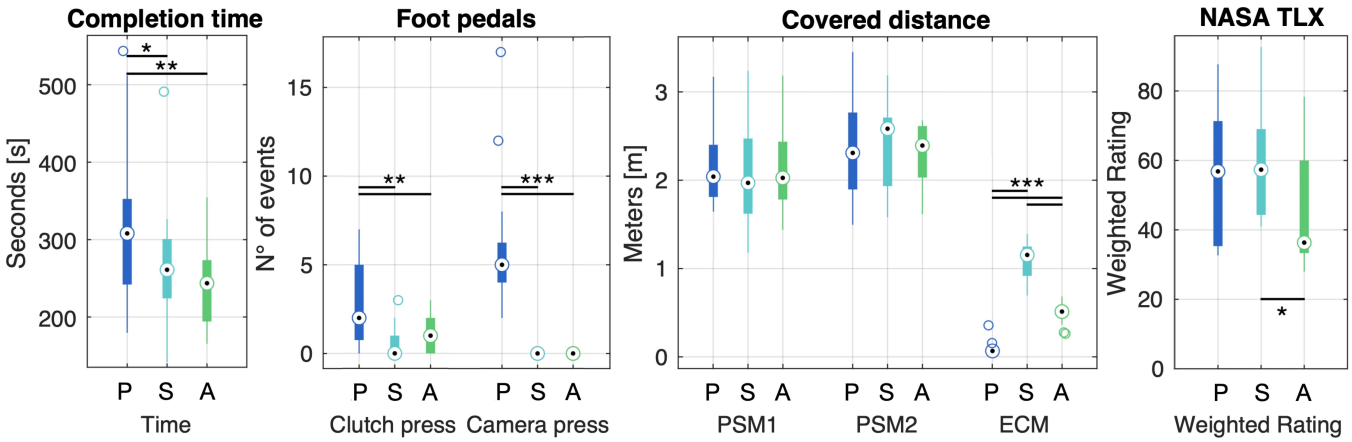


Fig. 3. Objective & subjective metrics: both objective (completion time, foot pedals, covered distance) and subjective (NASA TLX) metrics are shown. Groups are labeled based on camera motion modality, as follows: camera foot pedal (P), SCAN (S) [7], GRACE (A). Statistically significant differences are presented by pairs, identified by the extremes of the black horizontal lines. In the boxplots, the median is identified by a white-edged black dot, the first and the third quartiles are depicted as bold line edges, the whiskers are represented by thin line edges, while outliers are identified as rings.

both LSTM models 10 times, each time leaving out the labeled data coming from one user, obtaining 10 different versions. Every model was then used to perform online gesture recognition only for that specific user whose data was not used for training. Doing so, the gestures to be recognized given as input to the model always came from a never seen before user. Once the acquisition process was completed, the gestures' ground truth was obtained by offline manual labeling, and later on compared with the gesture recognition model output for validation. The same approach was later used for the user study, employing 10 user-specific models to perform online gesture recognition.

*4) Models Performance:* the evaluation of the LSTM models has been conducted through the comparison of the ground truth against the online classified gestures. Results were analysed based on macro-averaged F1 score and categorical accuracy. Macro-averaging allows for extension to multi-class classification - treating all classes equally - from a binary situation in which metrics are computed as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} \tag{3}$$

with True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Accuracy is limited in evaluating precision and recall information; consequently, the macro F1 score is employed.

*5) Comparative Metrics:* both objective and subjective metrics were defined to analyze the user study outcomes.

To evaluate users' performance from an objective point of view, we selected 6 quantifiable metrics: completion time, number of foot pedals presses (clutch and camera) and total covered distance for PSM1, PSM2 and ECM. The clutch pedal addresses how many times users had to readjust their masters' position due to a bad positioning of the surgical tools: whenever the clutch pedal is pressed, masters can be moved freely without causing any motion of the PSMs.

After completion of the study, every user was asked to complete a NASA Task Load Index (TLX) [30] for every camera control modality. The NASA TLX addresses the subjective workload, as an overall score based on a weighted average of rating on six sub-scales (mental demand, physical demand, temporal demand, performance, effort, frustration). Additionally, each user filled out a post experiment questionnaire to primarily inquire about motion sickness.

Given the relatively small sample size, we applied the Wilcoxon signed rank test to perform non-parametric statistical significance tests using the *signrank()* command in MATLAB. We considered repetitions with different camera control modalities to build two populations with paired observations of a certain metric, and statistically significant results were assessed at different values of $p$ as follows: * for $p < 0.05$, ** for $p < 0.01$, *** for $p < 0.001$. The results of statistical tests are reported as $p_{met}^{M-N}$, where *met* is the observed metric and $M$, $N$ are two camera control modalities.

## III. Results and Discussion

Every user who took part in the study completed the suturing task 2 times for each camera motion modality. The final aim of the study is to determine whether a situation aware autonomous camera control modality benefits the user during a suturing task, compared to both the current foot pedal control and a state of the art autonomous camera motion system. Results of gesture recognition coming from the user-specific models are shown in the confusion matrix in Fig. 4, and categorical accuracy and F1 score are reported. Both objective and subjective results are shown in Fig. 3.

### A. Models Performance

Results in Fig. 4 are obtained by calculating the confusion matrix of all 20 repetitions of modality A performed with user-specific models. With a similar approach, macro-averaged F1 score and categorical accuracy are 0.84. Errors in kinematics may affect the performance of the models as well as the tracking precision of the tool tips. In this context, it is noteworthy to report that the first generation of

the da Vinci robot may present position inaccuracies up to the order of a millimeter [31]. Nevertheless, automation of camera motion does not require sub-millimeter accuracy, in contrast to automation of surgical tasks involving grasping [31]. Therefore, we consider performance results sufficient for our application. Also, the vast majority of misclassified gestures fall at the boundaries of consecutive gestures. Such a result can be explained by two main factors: manual annotation, which may present inaccuracy, and latency. With the latter, we refer to the computational time required by the models to recognize the performed gesture and subsequently decide whether or not to move the camera. In fact, it took approximately $30ms$ to perform a single online gesture classification. This led to the system's working frequency reduction from $30Hz$ to $7Hz$; as a result, gesture progress at the boundary may be recognized with variable delay.

Consecutive gestures TB and NP show the higher misclassification values, as a result of their similarity in the execution, especially at the borders. Stronger results can be achieved by fusing multiple data streams. Incorporating video as input data would allow for estimation of needle pose and distance to tissue. Specifically, distance to tissue can be key in classifying NP and TB at the borders [32], where the LSTM model encounters challenges. NP and TB gestures get mixed in most cases due to their similarity during the transition from a fine position adjustment to find the correct stitch's insertion point (NP) to the start of the actual biting process (TB). This step witnesses minimal motion of the instruments, making the sole discriminative kinematic feature the orientation of the tools. Ultimately, incorrect detection of surgical gestures produces a sub-optimal FOV selection and camera instability, which result in mitigation of the beneficial effects introduced by situation awareness.

### B. Comparative Metrics

The results demonstrate statistical differences in completion time, foot pedal tray usage and camera covered distance. As expected, the introduction of an autonomous control for camera motion reduced significantly the completion time, with respect to pedal camera control, for both SCAN ($p^{S-P}_{time} < 0.05$) and GRACE ($p^{A-P}_{time} < 0.01$). GRACE improved completion time by a median reduction of $18.9s$ $(8.1\%)$ with respect to S and $65.1s$ $(21.1\%)$ with respect to P.

Regarding clutch pedal and camera pedal total presses, the same statistical difference is shown for both autonomous modalities, respectively $p^{A-P}_{clutch}, p^{S-P}_{clutch} < 0.01$ and $p^{A-P}_{camera}, p^{S-P}_{camera} < 0.001$. This illustrates the fact that during camera motion performed with the current traditional approach, hands are prone to fall into uncomfortable positions, which require the use of the clutch pedal for repositioning. In addition, the results can be explained by the autonomous nature of both SCAN and GRACE systems, which do not require any human input to directly control the camera, hence no camera pedal usage.

No reduction in the PSMs covered distance can be noted, but remarkable differences are shown for the ECMs covered distance. Even though for both autonomous navigation systems the covered distance is higher than the current camera control, we can notice a noteworthy statistical difference, with $p^{A-S}_{ECMdistance} < 0.001$, between GRACE (A) and SCAN (S). This result is given by two main factors: introduction of situation awareness and projection of camera center over the preferential line designated by the user at the beginning of every task. Indeed, online gesture recognition enhances camera motion by instructing the autonomous tracking system when to move the camera, while the projection of the point to be tracked over the line in 3D space further reduces unnecessary ECM movements, with the potential to reduce the incidence of motion sickness.

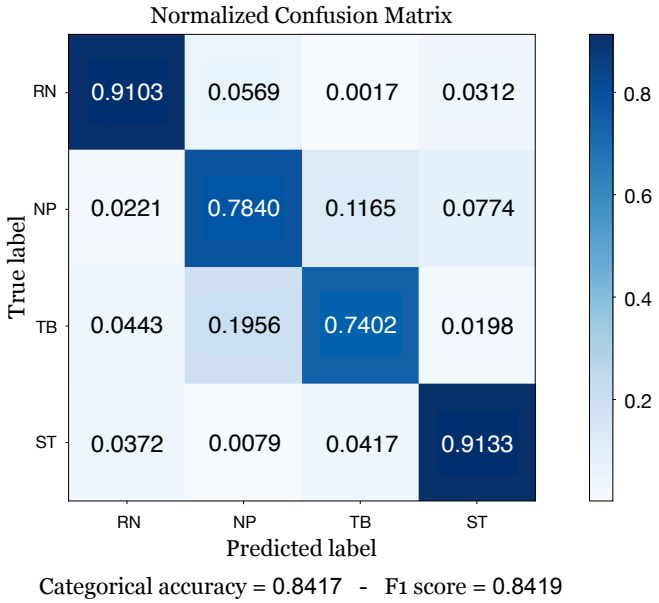


Fig. 4. Models Performance: gesture recognition results - obtained by user-specific models - are shown in the normalized confusion matrix, along with macro-averaged categorical accuracy and F1 score.

This suggests that automation in camera motion may not be sufficient to enhance the surgical outcome. In fact, a continuous motion of the camera, in particular when performing fine gestures such as biting the tissue, may be undesired, leading to motion sickness. The introduction of situation awareness with GRACE is able to drastically reduce unnecessary camera movements.

In order to complete the suturing task, the evaluated camera motion modalities required different levels of effort from the users. Such a result is confirmed by the NASA TLX subjective evaluation. A statistically significant difference between A and S demonstrates that GRACE is able to reduce the user's overall workload with respect to SCAN. However, no significant difference in terms of workload was found between A and S with respect to P. This result can be explained by the users' feedback in post experiment questionnaires. Indeed, results show that, while completing the task with modality P, users generally completed stitches even with sub-optimal camera FOV ($P = 3.5 \pm 0.85$ on a 0 - optimal - to 5 - suboptimal - scale). This suggests a decreased workload due to the fact that they partially skipped the effort to reposition the camera.

Furthermore, users reported a reduction of motion sickness

(*sickReduction*) using GRACE (A) compared to SCAN (S) ($A = 0.8 \pm 1.0$, $S = 3 \pm 1.6$ on a 0 - minimum - to 5 - maximum - motion sickness scale, with $p^{A-S}_{sickReduction} < 0.01$), confirming that situation awareness reduced the incidence of discomfort related to a continuous motion of the camera, typical of modality S.

## IV. Conclusions and Future Works

This paper presents an architecture to enhance camera navigation during a suturing task performed in robot-assisted surgery, thanks to the deployment of a system for online surgical Gesture Recognition for Autonomous Camera-motion Enhancement (GRACE). The system is based on two processes working together: online gesture recognition and kinematics-based tool tracking.

The system was integrated with the da Vinci Research Kit [29] and a user study was carried out to test its effectiveness with respect to state-of-the-art camera navigation. 10 subjects completed a suturing task in 3 camera motion modalities (manual, continuous tool tracking and GRACE). Results show that the proposed architecture is capable of reducing the burden associated to camera motion with respect to both the current control and state of the art autonomous tracking systems. Every user completed a post experiment questionnaire and results corroborate the hypothesis of motion sickness reduction thanks to the introduction of situation awareness, with respect to the state-of-the-art continuous non-aware tracking system, called SCAN [7]. Furthermore, reduction in overall workload with respect to SCAN demonstrates that situation awareness is a key factor in exploiting the beneficial effects brought by autonomous navigation. The fusion of an autonomous tracking system with an online model for gesture recognition further improved the overall surgical flow, not only relieving the surgeon from controlling the camera but also reducing its motion. Situation awareness allows for an intelligent motion of the camera by removing unnecessary movements that could cause motion sickness.

To improve the significance of the results, a larger population, including medical experts, should be involved in future studies. To further validate the method's generalizability and better transfer results from a dry lab to a real surgical scenario, uncertainties in the form of noise must be added, repeating also the experiments with $P = M$, when no preferential line is defined. In addition, further qualitative assessments of the surgical task should be performed. Also, the extension of gesture recognition to multiple surgical tasks is key towards of the general applicability of the proposed method. The fusion of both kinematic and video input data may result in higher performances for real-time gesture recognition. In particular, the extraction of meaningful features such as the distance of the needle from soft tissues, unattainable from kinematics alone, would reduce the uncertainty of the system in classifying gestures at the borders. To evaluate the model's ability to generalize, an evaluation of GRACE with models trained on single users' dataset should be carried out and compared to the presented results.

## References


[1] K. Moorthy, Y. Munz, A. Dosis, J. Hernandez, S. Martin, F. Bello, T. Rockall, and A. Darzi. Dexterity enhancement with robotic surgery. *Surgical Endoscopy and Other Interventional Techniques*, 18(5):790–795, 2004.

[2] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario. Medical robotics and computer-integrated surgery. In *Springer Handbook of Robotics*, pages 1657–1684. Springer, 2016.

[3] A. Pandya, L. A. Reisner, B. King, N. Lucas, A. Composto, M. Klein, and R. D. Ellis. A review of camera viewpoint automation in robotic and laparoscopic surgery. *Robotics*, 3(3):310–329, 2014.

[4] S. Ali, L. A. Reisner, B. King, A. Cao, G. Auner, M. Klein, and A. K. Pandya. Eye gaze tracking for endoscopic camera positioning: an application of a hardware/software interface developed to automate AESOP. *Studies in Health Technology and Informatics*, 132:4–7, 2008.

[5] D. W. Hansen, H. H. Skovsgaard, J. P. Hansen, and E. Møllenbach. Noise tolerant selection by gaze-controlled pan and zoom in 3D. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, pages 205–212, 2008.

[6] T. Da Col, G. Caccianiga, M. Catellani, A. Mariani, M. Ferro, G. Cordima, E. De Momi, G. Ferrigno, and O. De Cobelli. Automating endoscope motion in robotic surgery: A usability study on da Vinci-assisted ex vivo neobladder reconstruction. *Frontiers in Robotics and AI*, page 371, 2021.

[7] T. Da Col, A. Mariani, A. Deguet, A. Menciassi, P. Kazanzides, and E. De Momi. SCAN: System for Camera Autonomous Navigation in robotic-assisted surgery. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2996–3002. IEEE, 2020.

[8] B. Li, B. Lu, Z. Wang, F. Zhong, Q. Dou, and Y.-H. Liu. Learning laparoscope actions via video features for proactive robotic field-of-view control. *IEEE Robotics and Automation Letters*, 7(3):6653–6660, 2022.

[9] O. Weede, H. Mönnich, B. Müller, and H. Wörn. An intelligent and autonomous endoscopic guidance system for minimally invasive surgery. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5762–5768. IEEE, 2011.

[10] C. Gruijthuijsen, L. C. Garcia-Peraza-Herrera, G. Borghesan, D. Reynaerts, J. Deprest, S. Ourselin, T. Vercauteren, and E. Vander Poorten. Robotic endoscope control via autonomous instrument tracking. *Frontiers in Robotics and AI*, 9, 2022.

[11] A. D. Dragan, S. S. Srinivasa, and K. C. Lee. Teleoperation with intelligent and customizable interfaces. *Journal of Human-Robot Interaction*, 2(2):33–57, 2013.

[12] M. S. Yasar and H. Alemzadeh. Real-time context-aware detection of unsafe events in robot-assisted surgery. In *50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 385–397, 2020.

[13] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov. Gesture recognition in robotic surgery: a review. *IEEE Transactions on Biomedical Engineering*, 68(6), 2021.

[14] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, et al. Jhu-Isi Gesture and Skill Assessment Working set (JIGSAWS): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, 2014.

[15] S. S. Vedula, A. O. Malpani, L. Tao, G. Chen, Y. Gao, P. Poddar, N. Ahmidi, C. Paxton, R. Vidal, S. Khudanpur, et al. Analysis of the structure of surgical activity for a suturing and knot-tying task. *PloS One*, 11(3):e0149174, 2016.

[16] T. D. Nagy and T. Haidegger. A dVRK-based framework for surgical subtask automation. *Acta Polytechnica Hungarica*, pages 61–78, 2019.

[17] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

[18] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal. Sparse Hidden Markov Models for surgical gesture classification and skill evaluation. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 167–177. Springer, 2012.

[19] B. Varadarajan. *Learning and inference algorithms for dynamical system models of dextrous motion*. PhD dissertation, The Johns Hopkins University, 2011.

[20] L. Tao, L. Zappella, G. D. Hager, and R. Vidal. Surgical gesture segmentation and recognition. In *International Conference on Medical*

*Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 339–346. Springer, 2013.

[21] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pages 47–54. Springer, 2016.

[22] I. Funke, S. Bodenstedt, F. Oehme, F. v. Bechtolsheim, J. Weitz, and S. Speidel. Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 467–475. Springer, 2019.

[23] D. Sarikaya and P. Jannin. Towards generalizable surgical activity recognition using spatial temporal graph convolutional networks. *arXiv preprint arXiv:2001.03728*, 2020.

[24] R. DiPietro, N. Ahmidi, A. Malpani, M. Waldram, G. I. Lee, M. R. Lee, S. S. Vedula, and G. D. Hager. Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. *International Journal of Computer Assisted Radiology and Surgery*, 14(11):2005–2020, 2019.

[25] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov. Multi-task recurrent neural network for surgical gesture recognition and progress prediction. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1380–1386. IEEE, 2020.

[26] R. D. Ellis, A. J. Munaco, L. A. Reisner, M. D. Klein, A. M. Composto, A. K. Pandya, and B. W. King. Task analysis of laparoscopic camera control schemes. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 12(4):576–584, 2016.

[27] A. A. Awale and D. Sarikaya. Using human gaze for surgical activity recognition. In *2022 30th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE, 2022.

[28] N. Pasini, A. Mariani, A. Munawar, E. De Momi, and P. Kazanzides. A virtual suturing task: proof of concept for awareness in autonomous camera motion. In *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*, pages 376–382. IEEE, 2022.

[29] P. Kazanzides, Z. Chen, A. Deguet, G. S. Fischer, R. H. Taylor, and S. P. DiMaio. An open-source research kit for the da Vinci® Surgical System. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6434–6439. IEEE, 2014.

[30] S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier, 1988.

[31] Z. Cui, J. Cartucho, S. Giannarou, and F. Rodriguez y Baena. Caveats on the first-generation da Vinci Research Kit: latent technical constraints and essential calibrations. *arXiv e-prints*, pages arXiv–2210, 2022.

[32] C. Lea, G. D. Hager, and R. Vidal. An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 1123–1129. IEEE, 2015.