2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

An Intelligent Robotic Endoscope Control System Based on Fusing Natural Language Processing and Vision Models

Beili Dong*[1], Junhong Chen*[1], Zeyu Wang[1], Kaizhong Deng[1], Yiping Li[1]
Benny Lo[1], *Senior Member, IEEE*, George Mylonas[1]

***Abstract*— In recent years, the area of Robot-Assisted Minimally Invasive Surgery (RAMIS) is standing on the the verge of a new wave of innovations. However, autonomy in RAMIS is still in a primitive stage. Therefore, most surgeries still require manual control of the endoscope and the robotic instruments, resulting in surgeons needing to switch attention between performing surgical procedures and moving endoscope camera. Automation may reduce the complexity of surgical operations and consequently reduce the cognitive load on the surgeon while speeding up the surgical process. In this paper, a hybrid robotic endoscope control system based on fusion model of natural language processing (NLP) and modified YOLO-V8 vision model is proposed. This proposed system can analyze the current surgical workflow and generate logs to summarize the procedure for teaching and providing feedback to junior surgeons. The user study of this system indicated a significant reduction of the number of clutching actions and mean task time, which effectively enhanced the surgical training.**

I. INTRODUCTION

Robot-assisted minimally invasive surgery (RAMIS) benefits patients by reducing tissue trauma and hospitalisation time. Nowadays, the integration of Artificial Intelligence (AI) into surgical robotics systems to enable autonomy has become a prevailing topic and is well regarded to be the future of RAMIS [1]. However, the current available commercial surgical robotics are mainly adopting a master-slave control strategy to map the motions from the surgeon console to robotic instruments. Therefore, as the robot is operated under the direct control of the surgeon without any understanding of the surgical task and the surrounding environment, it cannot actively assist the surgeon. As a result, the commercial surgical robots are just tools with limited intelligence, only assisting surgeon in simple tasks (Level 1, robot assistance) [2]. Therefore, the aim of this project is to progress a surgical robot's level of autonomy, by introducing an intelligent model.

Progressing to higher levels on the autonomy scale, can be achieved from different aspects. There are three potential ways to implement intelligent assistance functions. Firstly, the capability to identify objects in the camera view can improve its understanding of surgical scenes. The development of object detection algorithms, such as YOLO-v8 [3], has led to highly accurate object detection methods, allowing detection of a target object and providing corresponding position information. Secondly, through capturing and interpreting the surgeon's instructions, the context of the operation can be estimated. Thus, it is necessary for the surgical robot to communicate with the surgeon to capture and confirm instructions. The development of Natural Language Processing (NLP) enables the understanding of human oral commands containing ambiguous instructions and the generation of corresponding answers by a transformer-based model [4]. Furthermore, these task-specific language instructions can be transferred directly into executable robot motions [5], [6]. In the field of surgical robotics, language and speech control also provide efficient assistance to operators [7], [8], [9]. Thus, voice instructions could be a promising method to help surgeons interact with the surgical robot. Thirdly, the system could identify any potential non-standard procedures and generate a logbook to inform and help the surgeon to improve their surgical skills.

This paper proposes a hybrid intelligence system for assisting surgeons in RAMIS. The system can autonomously or semi-autonomously control the pose of the endoscopic camera based on the surgical task and on oral instructions from the surgeon. The proposed system mainly consists of an object detection module, a language perception module, and a procedure perception module. The purpose of the computer vision (CV) based object detection module, is to determine the position of objects in the camera view. The role of the language perception module is to interpret the result-oriented oral instructions into camera control and feedback information.The procedure perception module can record feedback information from the robot as well as any identified potential non-standard operations during the procedure.

The main contributions of this work can be summarised as follows:

- Introducing an NLP model from progress-orientated to result-orientated command.
- An automatic system which can generate a surgical procedure log based on vision models.
- An autonomous perception model which can analyze surgical procedures and provide feedback.

II. METHODOLOGY

A. Overview

The system framework based on the Robot Operating System (ROS) is shown in Fig. 1. The voice assistant receives the vocal instructions and inputs them into a deep learning model composed of NLP & CV models to understand the intention of the surgeon. The details of the NLP & CV model will be introduced in the following subsections.

*These authors contributed to this paper equally.

[1]Hamlyn Centre for Robotic Surgery, Institute of Global Health Innovation, Imperial College London, Exhibition Road, London, SW7 2AZ, UK Corresponding email: george.mylonas@imperial.ac.uk

The NLP model transmits the intent analysis results to the cloud server using the Message Queuing Telemetry Transport (MQTT) protocol [10]. After that, the output labels of the fusion model are sent to the autonomous control strategy of the endoscopic camera manipulator (ECM). The details of the control strategy will be introduced in subsection D. Finally, by feeding the trajectory data from the patient side manipulator (PSM) into a Gaussian mixture model (GMM) and combining with NLP&CV model, the system can generate a procedural logbook of the surgical tasks.

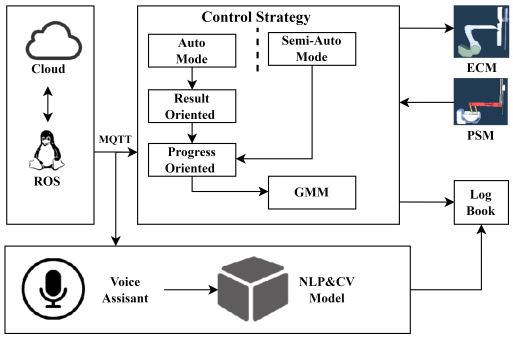

Fig. 1. The overview framework of the system

B. NLP Model

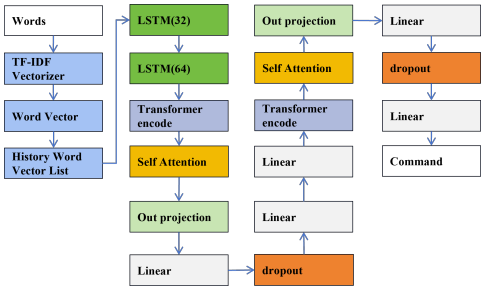

Fig. 2. The structure of the modified NLP Model

Fig. 2 shows the proposed NLP model. In the model, the Term Frequency - Inverse Document Frequency (TF-IDF) algorithm generates the word vectors from textual sentences [11]. In surgical contexts, dialogues and instructions often encompass interrelated steps, with subsequent commands referencing prior actions. Utilizing the capacity of Long Short-Term Memory (LSTM) model to process extended sequential data, NLP models can discern the intricacies and intentions of surgical commands. This allows the model to correlate a current instruction with earlier context, ensuring a consistent and precise interpretation of the surgical dialogue. The Transformer encoder layer (purple modules), comprising a self-attention (orange modules) and out-projection layer (light green modules), calculates positional relevance and provides weighted representations [12]. The self-attention mechanism calculates positional similarity, assigns weights, and thus enables superior context-dependent learning compared to LSTM [12]. Finally, a subsequent linear transformation in the out-projection layer remaps these dimensions to target dimensions, further enhancing feature extraction in dialogues [13].To mitigate the risk of over-fitting due to the limited corpus size, a dropout layer (red modules) is introduced as a regularization technique [14]. During each training iteration, this layer randomly deactivates neurons, thereby effectively training a new sub-network and increasing model adaptability [14].

In the modified NLP model, LSTM captures sequential information, while the Transformer assesses global information. The model synergizes the strengths of both cyclic and non-cyclic structures. In addition, when combined with the TF-IDF algorithm, the fusion model provides a nuanced understanding of dialogue context for automated ECM control. Compared with the ChatGPT model, the NLP model proposed in this paper would have better response performance on specific surgical tasks with relatively few computational resources and faster inferencing, which is expected to increase the efficiency.

C. CV Model

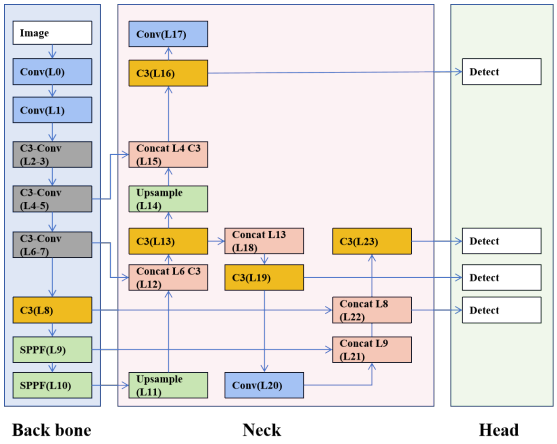

Fig. 3. The structure of modified CV Model

In order to follow the operating instruments in real-time, the proposed CV model was built on YOLO-V8 model [3] shown in Fig. 3. This modified YOLO-V8 model employs a Anchor-Free methodology, which replaces the conventional Anchor-Based technique. This approach directly predicts the spatial coordinates and dimensions of target object, eliminating the need for predefined anchors [15]. The Anchor-Free architecture specializes in calculating the geometric attributes of target, such as orientation and position, thereby augmenting the precision of target localization and potentially streamlining both the model's architecture and its training regimen [16].

The transition from the C2f to the C3 convolutional function occurs within the convolutional layer. Additionally, the Backbone section integrates the Spatial Pyramid Pooling Feature (SPPF) layer, facilitating the generation of multi-scale feature maps crucial for object detection tasks with targets of varying scales [16].

In the specific context of RAMIS, target objects are small in scale. To improve the small object detection performance, a Concat module is incorporated into the Neck section. This module enables the extraction of salient features from lower computational layers using a top-down semantic approach [17], thus improving the performance. Moreover, an auxiliary Detect output is created within the fully connected layer, merging high-level semantic features with low-level detail features. This composite structure enhances the semantic

feature set across various scales and improves the detection accuracy for small-size targets [18].

D. Control Strategy

1) ECM Automation Algorithm: This paper presents an ECM Automation Algorithm, shown in Fig. 4, that determines the focus point of ECM and scaling factor by integrating the CV model with kinematic data. The NLP model evaluates the field of view (FoV) requirements of the surgeon to either activate the autonomous tracking mode of ECM or locate tools outside the FoV. Subsequent sections discuss various ECM control modes, including autonomous and semi-autonomous options:

- **Semi-autonomous Camera (SAC) Control Strategy:** Users can trigger preset autonomous endoscope motion planning for subtle FoV adjustments using simple voice commands, such as: Move Left/Right, Zoom Out/In, Find Left/Right Tool, etc.
- **Autonomous Camera (AC) Control Strategy:** The system tracks the currently operated PSM and adjusts the position of endoscope and scaling factor to maintain FoV continuity and clarity. The focus center can be flexibly positioned to the left/right tool or to ensure the operated object is displayed near the centre region of interest area.

2) Algorithmic Assessment in AC Mode: In AC mode, the system categorizes the ongoing surgical operations into two types of procedures through algorithmic assessment:

- **Fine Motion:** When the displacement variation of PSM is minimal, indicative of precise operations like pattern cutting and knot tying.
- **Coarse Motion:** When the displacement variation of PSM is significant, indicative of faster, coarser operations.
- **Lifting:** Within Fine Motion, if the z-axis displacement significantly exceeds the x and y displacements, the action is categorized as a Lifting motion.
- **Moving:** Within Coarse Motion, if displacement of x or y axis significantly exceeds the z-axis, the action is categorized as Moving motion.

3) Fine-Tuning ECM Zoom Level and Viewpoint: After successfully categorizing the current PSM operation, the system fine-tunes the zoom level and viewpoint based on the identified motion type. Specific strategies are applied for Fine Motion, Lifting within Coarse Motion, and Moving within Coarse Motion to optimize the visual experience across different operational scenarios.

E. Procedure Perception Model

1) Repetitive Operation Detection: In conventional postoperative assessments, experts typically review extensive surgical footage to discern if certain repetitive actions during surgery are superfluous. Instead, the procedure perception system detects repetitive actions and produces a comprehensive procedure logbook, which includes all repetitive actions. This logbook is then presented to experts for verification of any superfluous actions. This approach considerably reduces both the postoperative evaluation duration and the workload of experts.

2) Data Collection and Analysis: During the surgical procedure, the system records kinematic data, which includes the position, speed, and direction of both the PSM and ECM. These data are subjected to cluster analysis using a GMM.

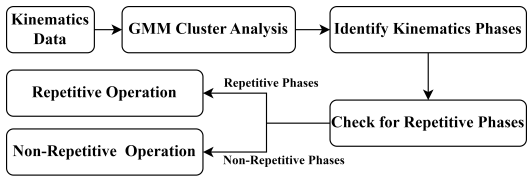

Fig. 5. The structure of Motion Phase Recognition Algorithm

3) Motion Phase Recognition: Motion phase recognition is used for repetitive operations identification. Since RAMIS often involves diverse motion patterns, the data naturally follows a multi-modal distribution. The multimodality of GMM enables effective identification of different motion patterns during surgery [19].

4) Logbook Analysis: The logbook summary system, as depicted in Fig. 5, employs GMM to analyze kinematic data. The GMM model excels in identifying patterns in the complex movements of sequential operations. Once the phase of a single-step operation is recognised, the system will display the ongoing operation phase. It can precisely record the operation type and detects its starting and ending points, along with nearby feature points. The system then evaluates whether the endpoint matches the targeted position. This guarantees comprehensive documentation of the nuances of operation, trajectory, and final destination for postoperative operation analysis.

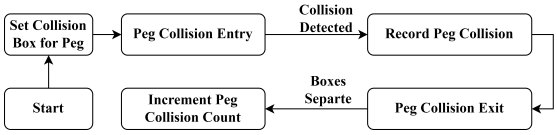

Fig. 6. The structure of Collision Algorithm

5) Collision Detection: The essence of algorithm is the precise (Fig. 6), real-time collision tracking between objects and instruments using specialized three-dimensional bounding boxes deployed on both columns and pegs. This feature enhances the adaptability of system, ensuring optimal performance even in rapidly changing and dynamic environments.

Collision Detection System:

- **Collision Entry:** Initiation of collision event recording and internal timer activation occur upon bounding box contact between pegs.
- **Collision Continuation:** Ongoing collision states are ascertained through continuous monitoring of relative positions and velocities.
- **Collision Exit:** The timer is halted and collision count incremented at the moment of bounding box separation.

6) Reviewing Module: When the system identifies a non-standard operation, the system categorizes the reviewing messages into two types: 'repeated-notice' and 'collision-notice'. These messages would be presented at the conclusion of the surgical logbook, enabling the surgeon

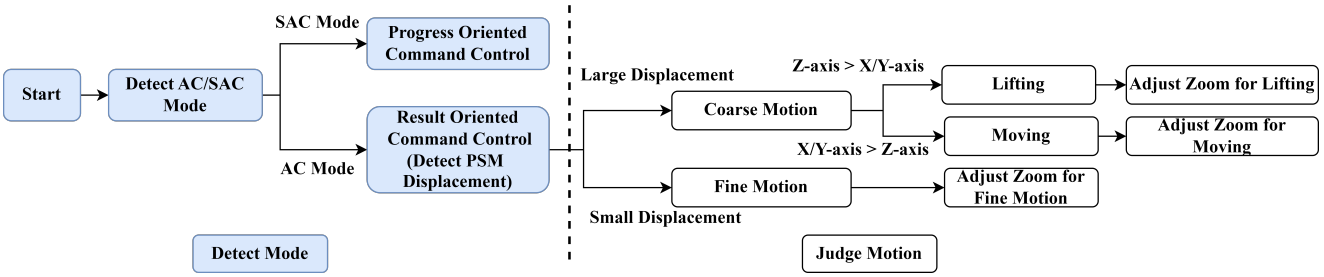

Fig. 4. The structure of ECM Automation Algorithm

to offer feedback upon task completion. These messages can inform the surgeons about the specific step where an issue may has occurred during the operation, effectively enhancing the training efficiency of surgeons.

III. Experiment Setup

A. Environmental Construction

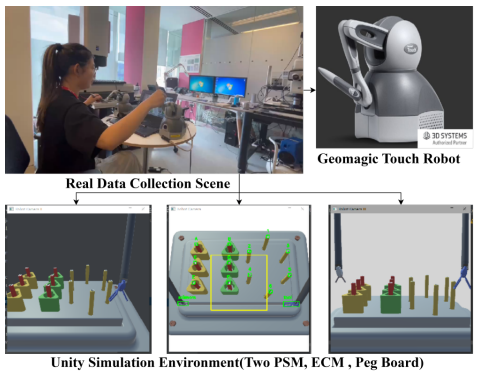

Fig. 7. The experiment setup

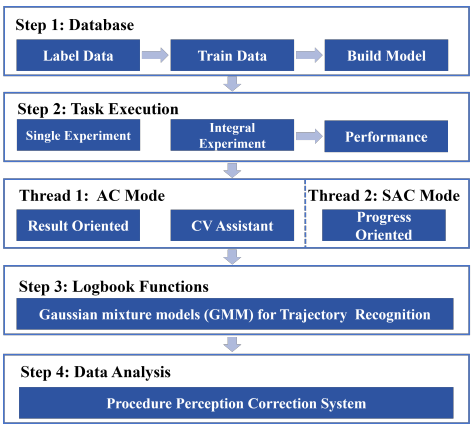

Fig. 8. The experiment flowchart

To validate the feasibility of the proposed system, the simulation environment of the da Vinci Surgical Robot (dVSR) built in [20] has been adopted in the experiment. As shown in Fig. 7, this simulation environment was built on UNITY (Unity Technologies, San Francisco, U.S.) to apply the designed model and implement the physical interaction features of each object. In the current experimental design, Peg Transfer was chosen to simulate a common surgical scenario, such as switching between fine and coarse operations during surgery. The peg transfer task was performed on a standard Fundamentals of Laparoscopic Surgery (FLS) training board [21], with twelve columns and six pegs on it. In addition to the training board, the simulation environment also included two PSMs placed on each side of the peg board and an ECM in the middle to provide the camera view. The main simulation window was set to the ECM viewpoint, where the operator gets the sight of the peg board and two PSM tools. Since the computer screen was not stereoscopic, a second camera view was adopted to provide the depth information of the experiment. As a generalized laparoscopic surgical training task, the da Vinci console cannot be used to effectively analyze the recorded data. Therefore, two Geomagic Touch haptic devices (3DSystems Inc., Rock Hill, SC, USA) were used as master manipulators for this experiment, to allow flexibility and adjustability to different experimental tasks.

A logbook generation module was built in the simulator to generate the corresponding task logbook after each experiment, which can be used to analyse and summarise the performance of a single task. The overall framework of the simulated procedure is shown in Fig. 8. The first step is to train and generate the proposed model by collecting experimental data from the simulated environment. The second step is to choose either SAC or AC mode to carry out the experiment, supplemented with CV and NLP model to perform the experiment and record the corresponding kinematic data. The third step is to generate a logbook by using GMM to analyze kinematics data and present all the experimental procedures. Then, in the last step, the procedure revision will list repetitive and collision operations.

B. Data Collection and Experimental Protocol

A total of nine subjects (8 males,1 female) participated in the experiment, all right-handed, between the ages of 22 and 30. Out of these, seven had experience with remote-controlled robots. Prior to the formal experiments, in order to eliminate the influence of the learning process of each subject on the results of the experiments, they were required to first perform a practice task to familiarise themselves with the ECM automation system. In the training phase, subjects were trained to use the AC and SAC modes. Subjects were not allowed to start the formal experiments until they had successfully completed 10 practice experiments. During the user study, 360 sets of kinematic data from the da Vinci simulated robot were collected for trajectory analysis and program logbook generation.

Differing from the baseline peg transfer experiment [22], in this work we require participants to place pegs at three

different target locations and return to the initial position at last. The initial position of PSMs were fixed to ensure identical initial conditions for unbiaced data collection and analysis. Kinematic data, including six degree-of-freedom (DoF) position information for two PSM tools, were recorded. As shown in Fig. 9, the peg transfer process was divided into 5 steps as follows:

a) Control the left surgical tool to pick up the peg from position E and place it at position 4 (left PSM tool).
b) Control the right surgical tool to pick up the peg from position 4 and place it at position 3 (right PSM tool).
c) Control the right surgical tool to pick up the peg from position 3 and transfer it mid-air to position D without placing it. (right PSM tool).
d) Control the left surgical tool to grasp the same peg at position D grasped by the right surgical tool, simultaneously transfer it mid-air at position C without placing (two PSM tools).
e) Move away the right surgical tool and control the left surgical tool placing the peg at position E (left PSM tool).

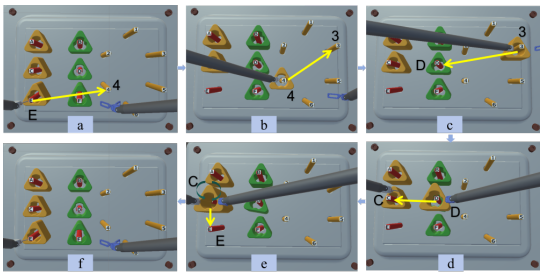

Fig. 9. The flowchart of modified experiment

C. Procedure Perception Experiment

An additional experiment based on the original task was designed to verify the effectiveness of the procedure perception model. In the experiment, subjects were asked to deliberately make non-standard operations to trigger this module. Non-standard operations were classified into two categories: repetitive operation and collision operation. Due to safety considerations of the actual cavity environment, the operations of the subjects were standardized within a certain range to ensure the robustness of the model. Six new subjects (5 males, 1 female) participated in this experiment to avoid over-fitting from original dataset. The participants, all right-handed, were aged between 22 and 30 years. And all of them had experience with remote-controlled robots. In order to validate the performance of the AC Mode, the experiment consisted of two tasks: an integral task and a single-step task. In the integral task, subjects were asked to complete the same peg transfer task as above, yet deliberately perform three random repetitive operations as well as three random collision operations in each round during the process. Upon completion of the task, the logbook was recorded and checked if it pointed out all deliberate non-standard operation. Each subject completed the integral task three times. In the single-step task, subjects were asked to deliberately perform both three random repetitive operations and three random collision operations in a single step. The logbook was analyzed once each single-step task was finished, checking if it had fully pointed out non-standard operations in each experiment. Subjects were required to perform the single-step task ten times per round, for a total of three rounds. In summary, data of 18 integral tasks and 180 single-step tasks were recorded.

IV. Results Analysis

In this section, a rigorous performance evaluation of the proposed system was conducted. This evaluation encompasses a baseline performance assessment for each individual component, namely the NLP Model, CV Model, and the procedure perception model. In addition, a user study has been conducted to empirically demonstrate the holistic effectiveness and outcomes of the proposed system.

A. Baseline performance assessment

1) NLP model assessment: In our experimental design, we divided the surgical audio dataset into 70% training set and 30% testing sets, and employed a 10-fold cross-validation methodology to assess its performance. Concurrently, we conducted a comparative analysis between the proposed Language Perception Model and one mainstream models: the conventional LSTM model [23].

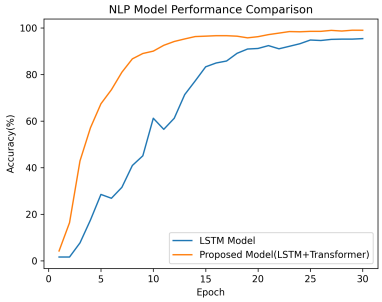

Fig. 10. The evaluation of NLP performance

As shown in Fig. 10, the proposed model exhibits a noticeably superior convergence rate and accuracy compared to the other models during the initial 15 training epochs. With an increase in the number of training epochs, the accuracy of system approaches 95.44%. Notably, owing to the incorporation of dropout and other strategic measures, the proposed model does not exhibit any discernible overfitting issues within the current dataset size. It is worth highlighting that, while the proposed model outperforms the LSTM model, their performance gradually converges with extended training epochs. In summary, the rapid convergence and high accuracy characteristics of the proposed model are promising to enhance the overall system performance, considering the potential future expansion of the dataset.

2) CV model performance assessment: In this section, we place a significant emphasis on evaluating the tracking accuracy of the CV model and the real-time performance within the signal propagation chain. As depicted in Fig. 11(a), we conducted a comparative analysis among several popular YOLO-V8 based CV models [3], considering both mAP@0.5, which effectively balances precision and recall, and mAP@0.5:0.95, which reflects the performance of model

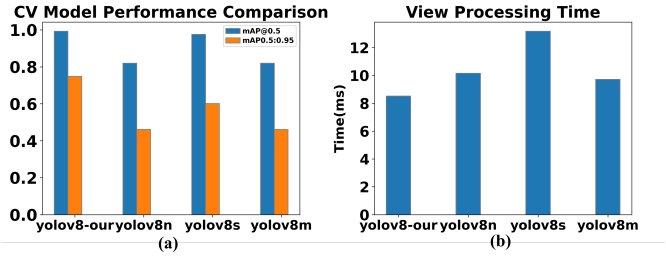

Fig. 11. (a) Comparison of CV Models (b) Evaluation of real-time performance

performance under stricter IoU thresholds. Experimental results unequivocally demonstrate that the proposed model exhibits optimal performance in both metrics.

Furthermore, in the evaluation of real-time performance, latency-related parameters primarily comprise three key components: pre-processing, inference, and post-processing, as illustrated in Fig. 11(b). It is noteworthy that the proposed model demonstrates the best overall performance among all comparative models, with a total latency of approximately 8.52 milliseconds.

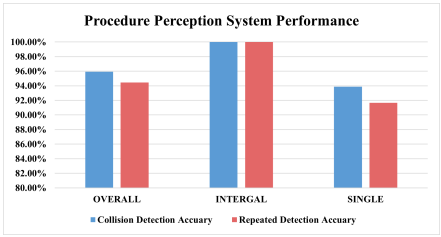

Fig. 12. The evaluation of Procedure Perception System

3) Procedure perception accuracy: The results of two experiments: an integral peg transfer experiment and a single-step experiment to simulate real-world scenarios, are depicted in Fig. 12. In the integral peg transfer experiment, both collision detection and repetitive detection achieved a remarkable accuracy rate of 100%. In the single-step experiment, the accuracy rates for these two aspects were 93.89% and 91.67%, respectively. On this basis, the overall accuracy is defined as a weighted average of the combined integral experiment accuracy and single-step accuracy. Taken together, the overall accuracy reaches 95.93%, 94.44%, signifying the effectiveness of the procedure perception model in evaluating surgical procedures.

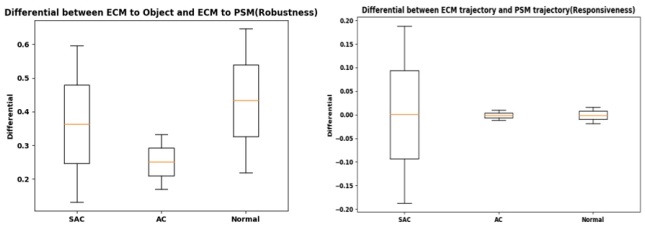

Fig. 13. (a) Robustness of System (b) Responsiveness of System

4) System performance evaluation: To evaluate the system performance, we compared the AC, SAC and Normal modes in terms of robustness and responsiveness.

Fig. 13(a) shows that the robustness of AC mode outperformed the others. This mode showed the least variability in view changes, indicating a more stable and focused performance with minimal outliers. In contrast, the SAC mode, while better than Normal, could not match the stability of AC mode due to its inability to distinguish between fine and coarse motions. While in Fig. 13(b) on responsiveness, the AC mode demonstrated superior performance. It maintained smaller variations in viewpoint movements over time, ensuring a more uniform speed and refined control. This coherence in the AC mode helps prevent issues like blurring or rapid scene changes, which are crucial for surgical precision.

B. User study results

After validating each individual components, the evaluation of the overall performance of the proposed system is presented in this section. As depicted in Table I, our user study checked the transfer efficiency, motion fluency, success rate, where MTT (Mean Transfer Time/seconds), TCT (Total Clutch Times/counts), and S/A (Success/Attempts) are defined respectively. It is clear that the AC mode outperforms the normal mode in terms of both MTT and TCT, while success rate also have a 2.7% increase comparing to the normal mode. This observation underscores the significance of the proposed system in assisting surgical procedures.

TABLE I

USER STUDY METRICS RESULTS

Control Mode	MTT-T(s)	TCT-avg(c)	S/A	S Rate(%)
AC Mode	111.7± 45.02	8± 6.4	287/300	95.7%
SAC Mode	118.9± 50.80	9± 6.8	284/300	94.6%
Normal Mode	116.4± 52.84	16± 6.8	279/300	93.0%

V. CONCLUSIONS

In this paper, an intelligent, robotic endoscope camera control system is proposed to assist robotic surgical operations. Three assisting models, namely NLP model, CV model, and procedure perception model, were developed in the system to identify surgical scenes, capture human-robot communication, and log the surgical procedure. The experimental results have shown the strength of these models. The results of the final user study indicate that task performance with this system outperforms task performance with only manual control. Regarding limitations, both the experimental design and simulation environment in this paper focus on a simplified peg transfer task to facilitate the validation of the system's functionality. Therefore, different parameter settings are required for other experiments, and task environments need to be partially replaced. If the system is to simultaneously address supervision and guidance functions for various tasks, separate system settings are necessary. Finally, since the CV model of this system is utilised in a single experimental environment, factors such as lighting conditions, task scales, and camera angles are not considered, which can be further validated in real surgical environments. Future work includes validation with a real surgical robot and qualified surgeons, construction of an operation guidance module, and real-time task kinematic and dynamic data analysis.

REFERENCES

[1] P. E. Dupont, B. J. Nelson, M. Goldfarb, B. Hannaford *et al.*, "A decade retrospective of medical robotics research from 2010 to 2020," *Science robotics*, vol. 6, no. 60, p. eabi8017, 2021.

[2] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, "Autonomy in surgical robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, pp. 651–679, 2021.

[3] G. Jocher, A. Chaurasia, and J. Qiu, "Yolo by ultralytics," https://github.com/ultralytics/ultralytics, 2023, accessed: 30th April, 2023.

[4] L. Ouyang, J. Wu, X. Jiang *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[5] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[6] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[7] M. Elazzazi, L. Jawad, M. Hilfi, and A. Pandya, "A natural language interface for an autonomous camera control system on the da vinci surgical robot," *Robotics*, vol. 11, no. 2, p. 40, 2022.

[8] Y. He, Z. Deng, and J. Zhang, "Design and voice-based control of a nasal endoscopic surgical robot," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 123–131, 2021.

[9] A. Rogowski, "Scenario-based programming of voice-controlled medical robotic systems," *Sensors*, vol. 22, no. 23, p. 9520, 2022.

[10] U. Hunkeler, H. L. Truong, and A. Stanford-Clark, "Mqtt-s-a publish/subscribe protocol for wireless sensor networks," in *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08)*. IEEE, 2008, pp. 791–798.

[11] B. Bhattarai, O.-C. Granmo, L. Jiao, R. Yadav, and J. Sharma, "Tsetlin machine embedding: Representing words using logical expressions," *arXiv preprint arXiv:2301.00709*, 2023.

[12] A. from the Chinese University of Hong Kong, "Simplified self-attention for transformer-based end-to-end speech recognition," 2023.

[13] R. Zhang, S. Frei, and P. L. Bartlett, "Trained transformers learn linear models in-context," 2023.

[14] Z. Liu, Z. Xu, J. Jin, Z. Shen, and T. Darrell, "Dropout reduces underfitting," 2023.

[15] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[17] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Lcnn: Low-level feature embedded cnn for salient object detection," 2015.

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[19] M. Edgington, Y. Kassahun, and F. Kirchner, "Dynamic motion modelling for legged robots," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 4688–4694.

[20] L. Qian, A. Deguet, and P. Kazanzides, "dvrk-xr: Mixed reality extension for da vinci research kit," in *Hamlyn Symposium on Medical Robotics*, 2019.

[21] G. Sroka, L. S. Feldman, M. C. Vassiliou, P. A. Kaneva, R. Fayez, and G. M. Fried, "Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial," *The American journal of surgery*, vol. 199, no. 1, pp. 115–120, 2010.

[22] J. Chen, D. Zhang, A. Munawar, R. Zhu, B. Lo *et al.*, "Supervised semi-autonomous control for surgical robot based on banoian optimization," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 2943–2949.

[23] K. Gopalakrishnan and F. M. Salem, "Sentiment analysis using simplified long short-term memory recurrent neural networks," *arXiv preprint arXiv:2005.03993*, 2020.