

Towards cognitive autonomous anatomy tracking in robotic surgery

Andrea Roberti¹, Daniele Meli², Marco Bombieri³, and Riccardo Muradore¹

Department of ¹Engineering for Innovation Medicine; ²Computer Science; ³Foreign Languages and Literatures
University of Verona, Italy

Autonomy is the frontier challenge of robotic surgery, with the potential to significantly improve the outcome of the surgical procedure, enhancing faster, precise motion, and the surgeons' experience, minimizing their fatigue and cognitive load. A crucial objective is the autonomous operation of the endoscope, in order to adjust the field of view of the surgical camera and focus on relevant parts of the scene. This can increase the attention and minimize the effort of the surgeon in robot-assisted surgery (complying to the requirement for human supervision imposed by the latest regulations on trustworthy artificial intelligence¹), and enhance perception capabilities of fully autonomous systems [1], [2]. Specifically, [2] evidenced the importance of tracking anatomies, instead of only the surgical instruments [1], in order to provide more task-specific guidance to the surgeon. On the other hand, they adopted a deep reinforcement learning model for autonomous camera motion, which is black-box and unpredictable, requires a large training dataset, and hardly generalizes out of its distribution [3]. In contrast, we propose a novel methodology for the cognitive autonomous tracking of *instruments and anatomies* in robotic-assisted surgery with the da Vinci Research Kit (dVRK). Following our experience in knowledge-based surgical task automation [4], we develop a pipeline to automatically retrieve a formal representation of the surgical procedure from text notes and standard descriptions [5]. This representation encodes the pre-conditions and effects of the surgical actions, together with their executing instruments and target anatomies. By monitoring the flow of pre-conditions during teleoperation (e.g., via automatic image segmentation and kinematic analysis), it is then possible to identify the ongoing surgical action and consequently use visual servoing to track its instrument and/or the target anatomy with the actuated endoscopic arm (ECM) of dVRK. In the context of simplified partial nephrectomy with phantoms, we compare our approach against teleoperation with fixed and human-operated camera. We show that our methodology improves the average accuracy in the recognition of the main anatomy, while minimizing the aggregate displacement of the instruments and slightly reducing the task duration.

I. METHODOLOGY

Our methodology is depicted in Figure 1. It starts with an *offline* phase for the *automatic extraction of task*

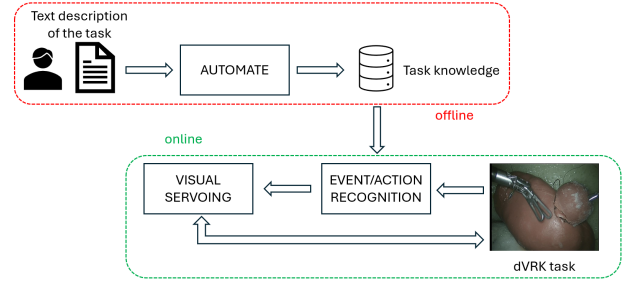


Fig. 1: Our framework.

knowledge from an available text describing the surgical task. The text satisfies minimal syntax rules, e.g., usage of the present or imperative tense. We then follow the approach of AUTOMATE [5] to automatically extract actions with pre-conditions and effects in a temporal formalism. Specifically, we use semantic role labeling (SRL) to identify verbs with subjects and main complements. The verbs (after univocal resolution of synonyms from WordNet [6], identify action names. Hence, we construct action predicates in the form $\text{action}(I, O)$, where I is the instrument (identified by SRL as a modality complement) and O is the object of the action, i.e., the target anatomy. For example, consider the sentence:

The surgeon dissects the tumor in a medial-to-lateral direction, using cold Endoshears.

Here, *dissect* is identified as a verb, hence action; *the tumor* as the object O ; *using Forceps* as the instrument I (by automatically filtering non-noun parts). With AUTOMATE, we also extract the *temporal sequence* of actions, as well as their pre-conditions and effects, by identifying temporal and conditional keywords in the text (e.g., *if, then, ...*). For instance, from the sentence:

Once the tumor is visible, the surgeon dissects the tumor in a medial-to-lateral direction, using cold Endoshears,

AUTOMATE identifies *the tumor is visible* as a pre-condition for $\text{dissect}(\text{Forceps}, \text{tumor})$.

In the *online phase*, we use the task knowledge to guide the automatic cognitive endoscope motion. Starting from the first action of the task, we identify its object or its instrument (the one with the highest recognition accuracy) as the target anatomy, and we use visual servoing to track it by moving the endoscopic arm, while the surgeon teleoperates the dVRK. Specifically, we use semantic segmentation [7] to pre-train a generative-adversarial network (GAN) efficiently on a very small set of surgical images;

¹EU Artificial Intelligence Act:
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>

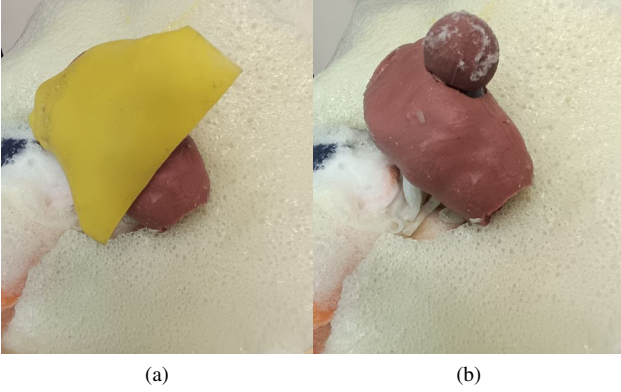


Fig. 2: Phantom with (a) and without fat (b).

we then identify the centroid of the target anatomy as the point to be tracked. In parallel, we also monitor the pre-conditions of the next action (e.g., the tumor visibility in the previous example) to trigger the event of action change.

II. RESULTS

We propose a preliminary evaluation of our methodology, in the context of the partial nephrectomy procedure with an anatomical phantom (Figure 2). The task consists of 3 main actions: removing the adipose tissue from the kidney defat(Prograsp,fat); dissecting the tumor dissect(Forceps,tumor); and suturing the wound suture(NeedleDriver,wound). Target anatomies/instruments are fat, tumor and NeedleDriver, respectively. For the semantic segmentation, we use the state-of-the-art Segment Anything Model (SAM) by Meta².

An expert dVRK operator performs 9 executions of the task, with the same initial ECM pose. Specifically, we compare 3 executions where the ECM is fixed (teleop), 3 executions where the operator moves the ECM towards the target anatomy (teleop move_cam), and 3 executions with autonomous ECM control (ours).

Figure 3a shows the tracking error (in pixels) of the target anatomy with the 3 algorithms. Moving the ECM (green) requires more execution time for the operator, while our autonomous control strategy (blue) is faster and more accurate. Figure 3b shows the segmentation (semantic) accuracy of recognition of the target anatomy. The highest accuracy is attained by our methodology, especially in the initial phase of the task (where fat is the target anatomy) and outperforming the teleoperation of the ECM when NeedleDriver is the dynamic (hence more challenging) target. These results are more evident in Table I, where the aggregate semantic accuracy is reported.

Our future research will focus on a more in-depth analysis of the impact of the proposed methodology on the dVRK operator’s experience, and validation across multiple more complex tasks.

²<https://segment-anything.com/>

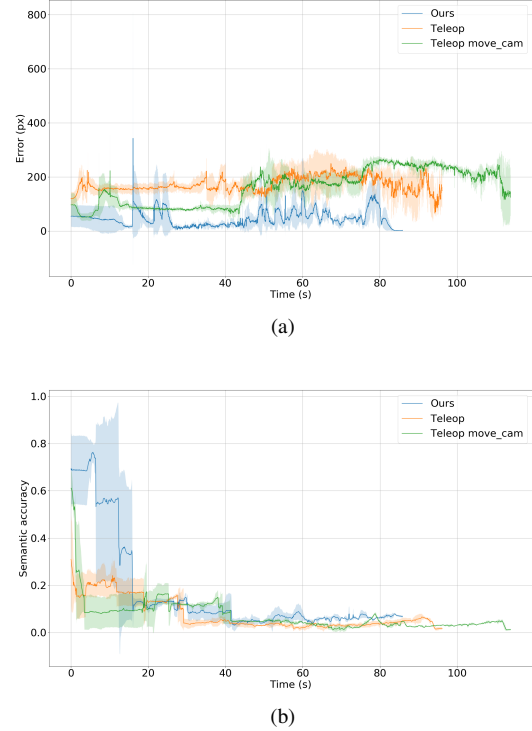


Fig. 3: Tracking error (top) and semantic accuracy (bottom). Mean $\pm$ std dev. across multiple executions.

Target	Ours	Teleop	Move_cam
fat	0.72 ± 0.15	0.24 ± 0.07	0.14 ± 0.14
tumor	0.10 ± 0.04	0.08 ± 0.05	0.07 ± 0.04
N.Driver	0.06 ± 0.02	0.03 ± 0.02	0.03 ± 0.01
Average	0.16 ± 0.22	0.08 ± 0.08	0.07 ± 0.08

TABLE I: Accuracy of the semantic segmentation (mean $\pm$ std deviation).

REFERENCES

- [1] Y. Huang, J. Li, X. Zhang, K. Xie, J. Li, Y. Liu, C. S. H. Ng, P. W. Y. Chiu, and Z. Li, “A surgeon preference-guided autonomous instrument tracking method with a robotic flexible endoscope based on dvrk platform,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2250–2257, 2022.
- [2] Z. Chen, K. Fan, L. Cruciani, M. Fontana, L. Muraglia, F. Ceci, L. Travaini, G. Ferrigno, and E. De Momi, “Toward human-out-of-the-loop endoscope navigation based on context awareness for enhanced autonomy in robotic surgery,” *IEEE Transactions on Medical Robotics and Bionics*, 2024.
- [3] R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel, “A survey of zero-shot generalisation in deep reinforcement learning,” *Journal of Artificial Intelligence Research*, vol. 76, pp. 201–264, 2023.
- [4] E. Tagliabue, D. Meli, D. Dall’Alba, and P. Fiorini, “Deliberation in autonomous robotic surgery: a framework for handling anatomical uncertainty,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 080–11 086.
- [5] M. Bombieri, D. Meli, D. Dall’Alba, M. Rospocher, and P. Fiorini, “Mapping natural language procedures descriptions to linear temporal logic templates: an application in the surgical robotic domain,” *Applied Intelligence*, vol. 53, no. 22, pp. 26 351–26 363, 2023.
- [6] C. Fellbaum, “Wordnet: An electronic lexical database,” 10 1998.
- [7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.