Semantic segmentation for tracking moving anatomies in robotic surgery*

Andrea Roberti[1], Daniele Meli[2], and Paolo Fiorini[1]

[1]*Department of Engineering for Innovation Medicine, University of Verona, Italy*

[2]*Department of Computer Science, University of Verona, Italy*

INTRODUCTION

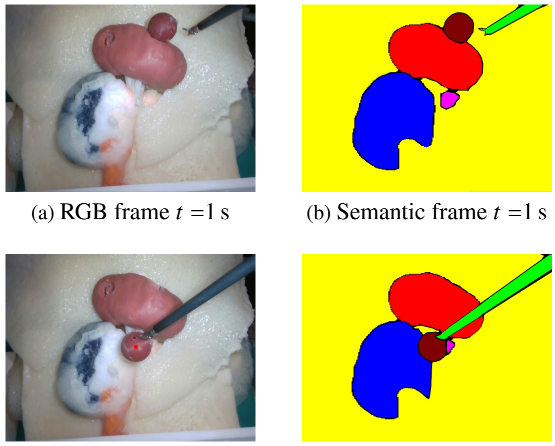
Robotics and Artificial Intelligence (AI) can bring invaluable benefits to clinicians in the surgical context. For instance, decision support systems and advanced autonomous systems for execution of standard repetitive sub-tasks will help reduce surgeon's fatigue, optimizing the outcome for the patient and minimizing usage of resources [1]. With the recent advances of machine learning and particularly deep learning, AI systems can achieve outstanding scene understanding [2], providing the surgeon with intra-operative guidance [3], e.g., via reality augmentation [4].

In the wide area of surgical robotic perception, one fundamental direction of research deals with the accurate tracking of laparoscopic tools for monitoring [5], skill assessment [6] and extraction of expert gestures for developing dexterous autonomous robotic systems [7]. However, several problems need still to be addressed. First, most deep learning-based systems require large datasets for robust training, as evidenced from the research boost to produce benchmarking datasets [8]. Moreover, instrument tracking is not enough to achieve proper situation assessment in surgery, but also accurate anatomical recognition is needed. One prominent data-efficient solution for this task is semantic segmentation [9], [10]. However, existing applications consider static or quasi-static scene, hence not accounting for moving anatomies due, e.g., to surgeon's manipulation and natural body motion.

In this paper, we address the problem of intra-operative anatomy tracking, using semantic segmentation trained on few RGB images of a realistic phantom for partial nephrectomy (Figures 1a-1c). We employ the da Vinci Research Kit (dVRK) in our experiments. We show that our methodology achieves comparable accuracy to the state of the art, over a wide motion range in the field of view of the RGB camera.

MATERIALS AND METHODS

We perform semantic segmentation using the Generative Adversarial Network (GAN) presented in [11]. It consists of a Generator (G) network based on U-NET [12] architecture, trained to produce multiple samples of a labelled image, starting from an input one. Another Discriminator (D) network is trained to model a loss function to identify whether the generated image is similar or not to the ground truth. The result is a pixel-to-pixel mapping model, where the G network is able to segment the elements in the scene, passing the test of the D network. We increase the robustness of the trained model, introducing a median and a clustering filter in order to reduce noise and scattered patches.



(a) RGB frame t =1 s (b) Semantic frame t =1 s

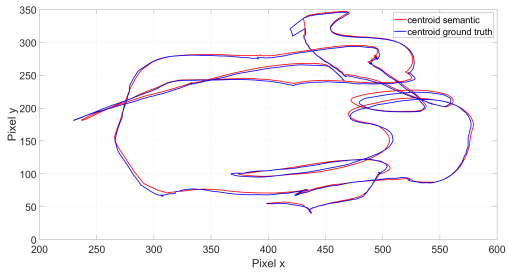
(c) RGB frame t =40 s (d) Semantic frame t =40 s

Fig. 1: The setup for partial nephrectomy with dVRK. On the left: RGB images from the endoscopic camera, with the centroid of the semantic blob of the tumor marked as red. On the right: semantic reconstruction.

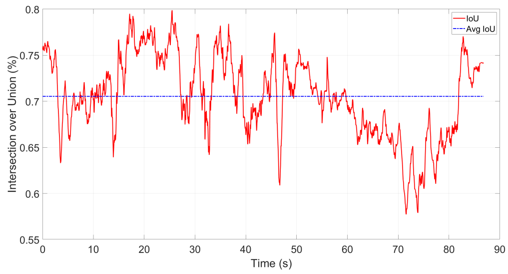
EXPERIMENTAL VALIDATION

We validate our model on images acquired through the 640×480 endoscopic camera of dVRK at 30 Hz, teleoperating the right tool to manipulate a tumor on the kidney surface. We preliminarily perform sub-millimetric hand-eye calibration following the methodology in [13]. Training requires manual annotation of only 10 frames (vs. thousands required, e.g., by U-NET), in order to label the features of interest, i.e., the tumor, the tools, the kidney and other anatomies. We obtain the qualitative results shown in Figure 1, for two example frames. We then perform a quantitative evaluation over the full record of execution, comparing to a ground-truth segmentation method based on standard color-shape segmentation in OpenCV. In Figure 2a, we show the trajectory performed by the centroids of the semantic blob of the tumor and of the ground truth, in pixel coordinates. We evaluate the Intersection over Union $IoU = \frac{|S \cap G|}{|S \cup G|}$, being S

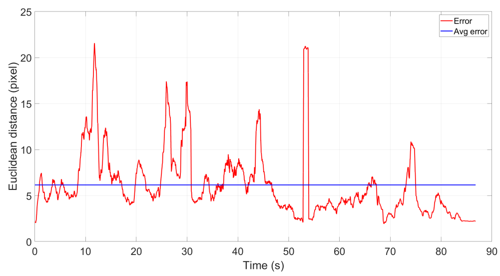
* This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No. 742671 (ARS).



(a) Centroid trajectory



(b) Intersection over Union



(c) Euclidean error

Fig. 2: Quantitiative results.

and G the set of pixels within semantic and ground-truth blobs and $|\cdot|$ the cardinality of a set; and the Euclidean error between the pixel trajectories. Results are shown in Figures 2b-2c, respectively. We achieve 0.72 ± 0.085 IoU over the full trajectory, which is comparable to state-of-the-art results presented, e.g., in [9] (≈ 0.74 for organs and ≈ 0.78 for tools). The Euclidean error is at most 20 pixel over a maximum of 640 pixel dimension of the camera, thus it is $< 3\%$ of camera resolution. The performance can be further improved adding some more frames in the training set, especially at the most critical points of trajectories (e.g., with the tool moving at high speed or with significant occlusion).

CONCLUSIONS AND DISCUSSION

In this work, we presented a data-efficient modified semantic segmentation architecture for robotic surgery, and validated it for tracking phantom anatomies from RGB images acquired with dVRK endoscopic camera. Compared to a state-of-the-art model-based segmentation methodology, our solution performs similarly to the state of the art in semantic segmentation of laparoscopic tools, achieving ≈ 0.72 IoU score. Moreover, it guarantees 20-pixel maximum Euclidean error over a 640×480 pixel resolution, thus achieving good accuracy for practical application.

In future research, we will integrate our technique into an autonomous active perception system for dVRK, with the ultimate goal to progressively refine anatomical awareness intra-operatively, improve localization of regions of interest and motion accuracy. This work represents a fundamental step towards autonomous and assisted surgical systems.

REFERENCES

[1] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, no. 4, p. 8638, 2017.

[2] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher, "Deep learning-enabled medical computer vision," *NPJ digital medicine*, vol. 4, no. 1, p. 5, 2021.

[3] R. B. Kochanski, J. M. Lombardi, J. L. Laratta, R. A. Lehman, and J. E. O'Toole, "Image-guided navigation and robotics in spine surgery," *Neurosurgery*, vol. 84, no. 6, pp. 1179–1189, 2019.

[4] J. Y. Chan, F. C. Holsinger, S. Liu, J. M. Sorger, M. Azizian, and R. K. Tsang, "Augmented reality for image guidance in transoral robotic surgery," *Journal of Robotic Surgery*, vol. 14, pp. 579–583, 2020.

[5] M. Islam, Y. Li, and H. Ren, "Learning where to look while tracking instruments in robot-assisted surgery," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V*. Springer, 2019, pp. 412–420.

[6] D. Lee, H. W. Yu, H. Kwon, H.-J. Kong, K. E. Lee, and H. C. Kim, "Evaluation of surgical skills during robotic surgery by deep learning-based multiple surgical instrument tracking in training and actual operations," *Journal of clinical medicine*, vol. 9, no. 6, p. 1964, 2020.

[7] M. Ginesi, D. Meli, A. Roberti, N. Sansonetto, and P. Fiorini, "Autonomous task planning and situation awareness in robotic surgery," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 3144–3150.

[8] J. Cartucho, S. Tukra, Y. Li, D. S. Elson, and S. Giannarou, "Visionblender: a tool to efficiently generate computer vision datasets for robotic surgery," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 9, no. 4, pp. 331–338, 2021.

[9] P. M. Scheikl, S. Laschewski, A. Kisilenko, T. Davitashvili, B. Müller, M. Capek, B. P. Müller-Stich, M. Wagner, and F. Mathis-Ullrich, "Deep learning for semantic segmentation of organs and tissues in laparoscopic surgery," *Current Directions in Biomedical Engineering*, vol. 6, no. 1, 2020.

[10] L. Tanzi, P. Piazzolla, F. Porpiglia, and E. Vezzetti, "Real-time deep learning semantic segmentation during intra-operative surgery for 3d augmented reality assistance," *International journal of computer assisted radiology and surgery*, vol. 16, no. 9, pp. 1435–1445, 2021.

[11] 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.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[13] A. Roberti, N. Piccinelli, D. Meli, R. Muradore, and P. Fiorini, "Improving rigid 3-d calibration for robotic surgery," *IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 4, pp. 569–573, 2020.