A Multi-Modal Learning System for Action Segmentation to Control Assistant Surgeon Robots

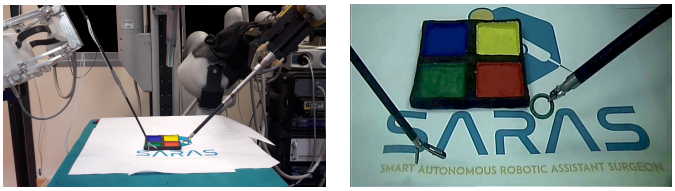
Giacomo De Rossi, Serena Roin, Fabio Falezza, Francesco Setti, Riccardo Muradore
University of Verona
Italy
name.surname@univr.it

***Abstract*—The next stage for robotics development is to introduce autonomy and cooperation with human agents in tasks that require high levels of precision and/or that exert considerable physical strain. To guarantee the highest possible safety standards, the best approach is to devise a deterministic automaton that performs identically for each operation. Clearly, such approach inevitably fails to adapt itself to changing environments or different human companions. In a surgical scenario, the highest variability happens for the timing of different actions performed within the same phases. This paper presents a multi-modal action segmentation system that operates online to specifically target semi-autonomous assistive robotic platforms during Robotic-Minimally Invasive Surgery (R-MIS). It provides the required timing for the actions being performed to direct the lower-level control of the assistant robot through supervisory control.**

***Index Terms*—Action Recognition, R-MIS, Shared-Control Systems**

I. Introduction

The autonomous execution of a task by robots is mostly relegated to industrial applications where robotic platforms execute repetitive tasks with minimal to no cooperation with humans: the focus is on executing precisely the same motions in the most efficient way when positioned in a highly structured environment. The research in robotics, however, is pushing for the introduction of cooperative tasks in which both the motion accuracy and cognition level need to be robust under any condition [1]. In medical robotics, the main effort is nowadays spent in the development of autonomous and semi-autonomous technologies to R-MIS. A comprehensive study performed in [2] evaluates the impact of autonomous technologies on medical/surgical practice and emphasises the need of human cooperation and supervision in the future of autonomous robotic surgeries. Among many applications available in literature, the most relevant ones are the recognition of the different phases in an endoscopic surgery addressed with deep neural networks [3] and the implementation of a knowledge-based ontology approach [4]. The SARAS[1] SOLO-SURGERY platform will be a very sophisticated example of a shared-control system: a surgeon operates remotely a pair of robotic laparoscopic tools (i.e. the daVinci® Surgical Platform) and cooperates with the two novel SARAS autonomous robotic arms inside a shared environment to perform complex surgical procedures. The goal of the project is to substitute the assistant surgeon next to the patient within the operating room. This paper showcases an approach to deduce human actions from the robotic minimally-invasive surgical scene to coordinate a semi-autonomous robot with a human surgeon. Its main contributions over the state-of-the-art is represented by a multi-modal cognitive system for *action segmentation* that encompasses both visual and kinematic information with the adoption of a Temporal Convolutional Neural Network, which allows for a better identification of actions of different duration when compared against similar solutions.



(a) (b)

Fig. 1. Experimental setup. (a) daVinci® arm (right) and SARAS arm developed by Medineering™ (left). (b) same scene seen through the left endoscope camera.

II. Action Segmentation

The *action segmentation* module has to operate within stringent timing and performance requirements to be applied online as a soft-sensor. Indeed, the underlying model must:

- be reliable, which can be verified by the low incidence of false positives and negatives, and the percentage of correctly evaluated sequences;
- be robust, which is tested under varying conditions for the experimental setup (lighting, camera orientation, target variation etc.)
- provide real-time evaluation for its application as an advanced soft-sensor taking as input fast-changing signals and providing as output commands to lower-level controllers. This requires both data buffering operations and a small memory footprint not to hinder cyclic computations.

To comply with these requirements, we chose to implement a neural network, called *EdSkResNet* (Encoder-Decoder Spatial-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 779813 (*SARAS*). The authors would like to thank Johann Wigger and Sabine Hertle at Medineering GmbH for the development of the SARAS robots used in the experiments.

[1]SARAS is an EU founded project and stands for Smart Autonomous Robotic Assistant Surgeon, details at www.saras-project.eu

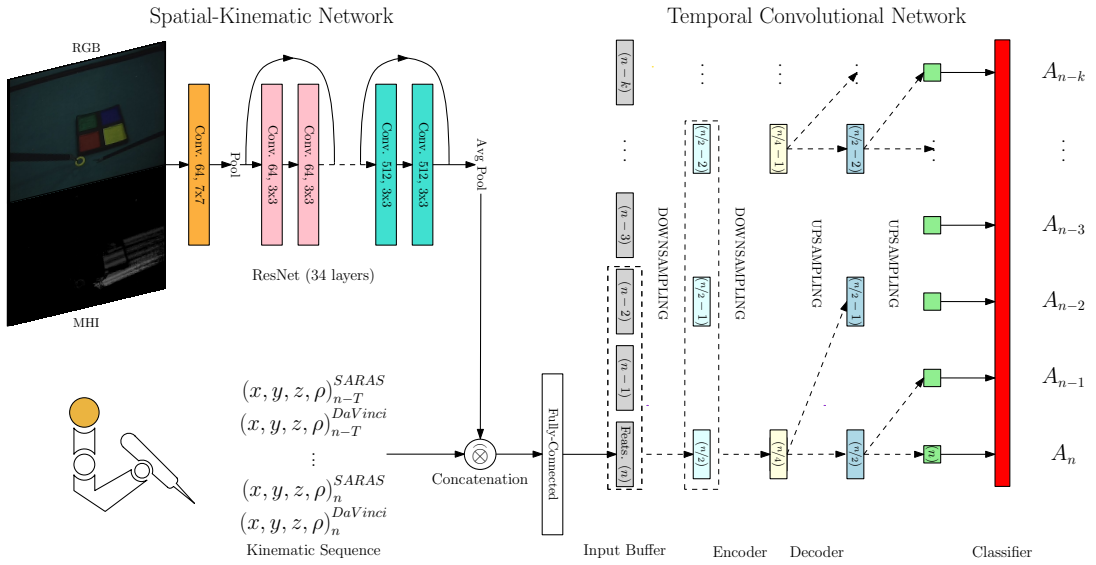


Fig. 2. Neural network schema for action segmentation: the RGB and MHI images are processed simultaneously as a 4-channel enhanced frame.

Kinematic ResNet) (Figure 2), that integrates multi-modal learning over the data available during a robotic minimally-invasive surgical operation. It outperforms many state-of-the-art solutions for both spatial, i.e. the instantaneous description of the scene, and temporal information analysis. The network is split in two halves: the *Spatial-Kinematic Network* analyses the spatial relation of all objects in the scene through RGB images, Motion History Images (MHI, the difference of the frames in a time window) and kinematic trajectories of the robots; the *Temporal Convolution Network* processes a temporal filter of the intermediate processed features to identify the various action segments. The main feature of this network is its causal formulation as the processed features are inserted in a circular input buffer and filtered through causal convolution operators. Additionally, the fully-convolutional structure improves both training and execution performance when compared against recurrent neural network formulations [5].

III. Results

Validation of the action segmentation has been performed on the SARAS setup over a custom surgical pick-and-place training task performed by an human operator teleoperating the daVinci® and the autonomous SARAS arm. The user is instructed to pick up a colored ring placed in the scene, either red, blue or green, and to bring it closer to the camera for color identification. The robot arm, using both cognitive and geometrical information inferred from image and kinematic data, moves towards the ring; after grasping it, the robot waits until the other arm releases the ring and, finally, leaves the exchange area to deliver the ring to the corresponding target by color. The action segmentation identifies 8 actions being performed by the surgeon and the robot and drives the controller for the latter. The final control action can be seen in the attached video. An additional validation is provided in Table I with a comparison of the result of the *EdSkResNet* with the best-performing algorithm in literature [6] for the JIGSAWS [7] dataset. The proposed methods improves the Accuracy (percentage of correct actions) and Edit Score (number of correct sequences, without repetition) indicate the prowess of the model to identify the correct sequence of actions. It needs to be specified that the method in [6] operates in a non-causal manner as it performs non-causal temporal filtering.

TABLE I
Results for the Median User of the JIGSAWS Suturing (%).

Algorithm	Accuracy	Edit Score
ED-TCN [6]	81.4	83.1
EdSkResNet	**81.71**	**91.74**

References

[1] M. Bonfe, F. Boriero, R. Dodi, P. Fiorini, A. Morandi, R. Muradore, L. Pasquale, A. Sanna, and C. Secchi, "Towards automated surgical robotics: A requirements engineering approach," in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2012, pp. 56–61.

[2] F. Ficuciello, G. Tamburrini, A. Arezzo, L. Villani, and B. Siciliano, "Autonomy in surgical robots and its meaningful human control," *Paladyn, Journal of Behavioral Robotics*, vol. 10, no. 1, pp. 30–43, 2019.

[3] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[4] O. Dergachyova, X. Morandi, and P. Jannin, "Knowledge transfer for surgical activity prediction," *International journal of computer assisted radiology and surgery*, vol. 13, pp. 1409–1417, 2018.

[5] E. Tsironi, P. Barros, C. Weber, and S. Wermter, "An analysis of Convolutional Long Short-Term Memory Recurrent Neural Networks for gesture recognition," *Neurocomputing*, vol. 268, pp. 76–86, 2017.

[6] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9915 LNCS, pp. 47–54, 2016. [Online]. Available: http://arxiv.org/abs/1611.05267

[7] 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*, vol. 3, 2014, p. 3.