IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Autonomous Tissue Retraction in Robotic Assisted Minimally Invasive Surgery – A Feasibility Study

Aleks Attanasio,[1] Bruno Scaglioni[1], Matteo Leonetti[2], Alejandro F. Frangi[2],
William Cross[3], Chandra Shekhar Biyani[3], Pietro Valdastri[1]

***Abstract*— In this work, we describe a novel framework for planning and executing semi-autonomous tissue retraction in minimally invasive robotic surgery. The approach is aimed at removing tissue flaps or connective tissue from the surgical area autonomously, thus exposing the underlying anatomical structures. First, a deep neural network is used to analyse the endoscopic image and detect candidate tissue flaps obstructing the surgical field. A procedural algorithm for planning and executing the retraction gesture is then developed from extended discussions with clinicians. Experimental validation, carried out on a DaVinci Research Kit, shows an average 25% increase of the visible background after retraction. Another significant contribution of this paper is a dataset containing 1,080 labelled surgical stereo images and the associated depth maps, representing tissue flaps in different scenarios. The work described in this paper is a fundamental step towards the autonomous execution of tissue retraction, and the first example of simultaneous use of deep learning and procedural algorithms. The same framework could be applied to a wide range of autonomous tasks, such as debridement and placement of laparoscopic clips.**

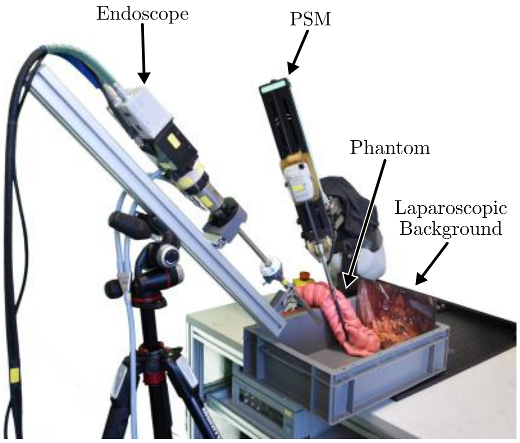

Fig. 1. DVRK setup composed of a PSM and a stereo endoscope. A phantom and a printed laparoscopic background have been used to validate the semi-retraction approach.

I. INTRODUCTION

Minimally Invasive Surgery (MIS) presents several benefits for patients compared to open surgery, such as reduced trauma to the anatomical structures, shorter recovery time, and reduced blood loss [1]. A significant portion of each MIS procedure is devoted to Tissue Retraction (TR), which is conducted to access the area of interest (e.g. tumour) [2]. Exposing the surgical area is therefore a crucial task in MIS, as surgeons rely mainly on visual information, given that tactile feedback is absent or extremely limited. This is especially problematic in urology, where access to the bladder and prostate is obstructed by bowels and connective tissue [2]. In this clinical practice, robotic MIS is nowadays a common approach, with platforms such as the DaVinci Surgical System (DVSS) from Intuitive Surgical widely used worldwide. The DVSS is a master-slave teleoperated system, i.e. the movements of the surgeon on two Master Tool Manipulators (MTM) are replicated on the tip of laparoscopic instruments by means of three Patient Side Manipulators (PSM). During a typical robotic MIS procedure, the surgeon temporarily assigns one of the MTMs to the third PSM to perform tissue retraction, or requires the support of an assistant to carry out the task with an additional manual instrument. Retraction often involves manipulation of connective tissues or organs (e.g., liver or bowel). Switching robotic arms, or instructing an assistant on the desired retraction motion, significantly increases the surgeon's cognitive load [3] and raises severe risks with potentially catastrophic consequences [4]. TR can also be challenging in the context of manual laparoscopy, where the lack of coordination between surgeon and assistant can lead to hazardous situations, such as instruments collisions, tissue damage or unintentional tearing [5]. To tackle these issues, this paper presents a semi-autonomous system for TR that can be applied to surgical procedures using a robot-controlled instrument (i.e., full robotic MIS or hybrid manual-robotic procedures).

Our approach focuses on detecting tissue flaps obstructing the surgical field by using U-Net [6], a particular convolutional neural network structure, widely adopted in the segmentation of medical images. The network (henceforth: U-Net), fed with the endoscopic video stream, is trained via a dataset of surgical images recorded during procedures performed on Thiel-embalmed cadavers (i.e. an embalming technique that preserves the softness of human tissues [7]), and subsequently labelled manually. An algorithm is developed to identify the retraction grasping point and direction based on the size and shape of the detected flaps. This enables the TR to be planned and then planned and performed autonomously.

[1] A. Attanasio, B. Scaglioni and P. Valdastri are with the Storm Lab UK, School of Electronic and Electrical Engineering, University of Leeds, Leeds, UK,{elaat,b.scaglioni,p.valdastri}[at]leeds.ac.uk

[2] M. Leonetti and A. F. Frangi are with School of Computing, University of Leeds, Leeds, UK, {m.leonetti,a.frangi}[at]leeds.ac.uk

[3] W. Cross and C.S. Biyani are with Department of Urology, St James University Hospital, Leeds, UK, {shekhar.biyani,williamcross}[at]nhs.net

This methodology was validated on a DaVinci Research Kit (DVRK) [8] and experiments were performed on a benchtop platform. However, the proposed approach could be applied to any other surgical MIS platform fitted with stereo vision and at least one instrument manipulated by a robot [9].

Research in surgical robotics has recently focused on increasing the level of robots' autonomy, with examples of automating tasks such as tool detection [10], suturing [11], and resection [12]. The research on task autonomy aims at relieving the surgeon of manual and repetitive tasks in a collaborative framework, rather than substituting the human action completely [13], [14], [15]. Research in autonomous suturing and related sub-tasks, discussed in [16], [17] has been greatly facilitated by the availability of datasets dedicated to the analysis and automation of surgical gestures (JIGSAWS [18]). The use of automation for 3D tissue debridement of soft tissues presented in [19] is particularly interesting. The literature on TR is limited, despite this task being repeatedly performed during all typical procedures. In [20] and [21], two simulation frameworks to perform a grasp-and-retract task are presented. In [20], a constant tissue curvature trajectory is proposed to minimise mechanical stress. Similarly, in [21], path planning for retraction in the presence of an obstacle is reported. More recently, advanced approaches have been proposed in [22], where retraction is controlled by an image-based system, and in [23], where three different approaches based on proportional control, hidden Markov models and fuzzy logic are developed. In these works, the start and end points of the retraction are manually indicated by the surgeon thus entailing no autonomous planning. Concerning the use of deep learning algorithms in the context of surgical data, the U-Net neural network has been developed for segmentation of biomedical images [6], and subsequently widely adopted in various surgical scenarios such as brain tumour detection [24], liver tumour tracking [25], and surgical tool detection [26]. In [27], segmentation is performed on MRI images, aiming at localising tumours by means of 3D reconstruction. However, U-net has not yet been applied to the detection of tissue flaps for the automation of retraction.

The main contribution of this work is a framework for semi-autonomous tissue retraction, including endoscopic image analysis and gesture planning. This contribution advances the field of robotic-assisted MIS, laying the foundations for future developments in the field of autonomous surgical assistance. Compared to other works in soft tissue retraction, such as [22] and [23], we increase the level of autonomy by providing autonomous tissue segmentation and gesture planning abilities directly on the endoscopic video sequence. Our system is capable of automatically extracting start and end points for tissue retraction, thus reducing the input required from the surgeon in defining task specifications. Other works, such as [15], adopt a similar workflow but focus on a different task (i.e. debridement) and therefore develop algorithms specifically dedicated to debris detection. Another major contribution of the present work is the introduction of FlapNet, a dataset of labelled surgical images dedicated to retraction, available at https://github.com/Stormlabuk/FlapNet. The dataset offers a valuable resource for research in the field of anatomy navigation. The approach described here leverages both deep learning techniques, well-suited to image analysis, and procedural algorithms, which offer the advantage of predictable behaviour and repeatability. The same approach can be adopted to perform other semi-autonomous tasks such as ablation, placement of laparoscopic clips, and debridement.

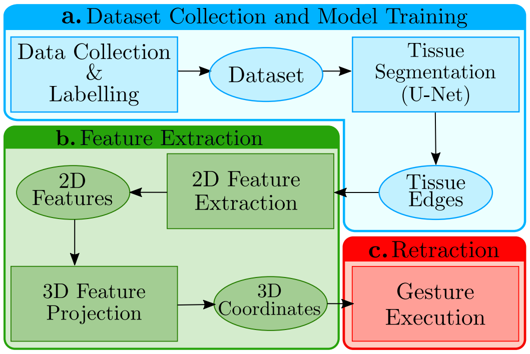

Fig. 2. Tissue retraction pipeline: a U-Net is trained using manually labelled disparity maps evaluated from stereo images of a cadaveric lobectomy. Subsequently, 2D features such as grasping point, background and tissue centroids are identified on the tissue mask output by the network. Finally, the features are projected in the 3D space by means of epipolar geometry, allowing the DVRK controller to plan and perform the retraction.

II. MATERIALS AND METHODS

In Figure 2, a schematic diagram of the proposed method is represented. The approach is composed of three main elements: Tissue flaps detection (Fig. 2-a), extraction of relevant features (Fig. 2-b), and gesture planning and execution (Fig. 2-c). The output of each stage corresponds to the input of the following stage. In this work, a "detect-plan-execute" approach is adopted to allow the surgeon to maintain control over the execution of the gestures. The system is designed to plan the retraction and subsequently show the surgeon the grasping point, the retraction direction and the final position of the tool. The surgeon can acknowledge the execution by means of a pedal or voice control. The retraction gesture is performed for as long as the surgeon maintains pressure on the pedal. To avoid loss of visual control on the instrument, the camera field of view is mapped on the workspace, and motion of the tool is limited within the image's boundaries, whereby the boundaries correspond to the full-size image cropped by 5%.

A. Tissue Flap Detection

The initial stage of the retraction process is the detection of the tissue flap to be retracted. This feature is provided by a U-Net developed in the Tensorflow [28] framework. The network is characterised by 5 encoder and decoder blocks. Each encoder, composed of 2 convolutional layers with batch normalisation and a Rectified Linear Unit (ReLU) acting as activation function, outputs into a max pooling layer with pool size 2. The decoder is composed of 3 convolutional layers with batch normalisation and ReLU activation function and the feature map is expanded by a factor of 2. The output is a convolutional layer with sigmoid activation function and 1 neuron. In order to avoid overfitting, dropout is applied to the 3 encoders and decoders closer to the centre of the network. Starting from the first encoder, which includes 32 units, the following encoders are characterised by an increasing number of neurons (i.e. double at every step), to reach a maximum of 1,024 at the centre of the network. Conversely, the number of units per encoder is decreased by a factor of 2 moving from the centre to the output layer. In order to enhance robustness with respect to different anatomical structures and colours, RGB depth maps (DM) are adopted as input for the neural network. DMs are images [29] in which the intensity of every pixel is associated to a defined distance from the camera lens. In this work, DMs are created base on the disparity between left and right images produced by the DVSS stereo camera. For this reason, they are robust to changes in lighting conditions and tissue reflections. Moreover, DMs are colour-blind, thus not varying based on the colour of different organs and tissues. As the goal of the U-Net is to detect the candidate flaps for retraction, a grayscale mask of the same size of the input DM is chosen as output, where the value of each pixel, from 0 to 1, describes the likelihood of a tissue flap appearing in that pixel. An example is shown in Figure 3.

B. Dataset Collection

In order to create a training dataset for the U-Net, video streams of surgical procedures (lobectomy) performed on a single Thiel-embalmed cadaver by experienced surgeons using a DaVinci Xi have been collected. Starting from stereo image pairs (i.e., left and right cameras), DMs are generated by means of the `stereo_img_proc` ROS package, which is based on a modified version of the Semi-Global Matching algorithm [30], available in OpenCV [31]. Under the assumption of consecutive images being very similar, as the movement of the camera is slow and discontinuous, a three-steps approach is adopted to maximise the variability between images before manual labelling.

- The 356 minutes long video file of the procedure is reduced manually to 62 minutes by selecting the most relevant parts of the procedure where one or more retraction is performed.
- One pair of images is sampled every second, resulting in a set of 3,720 pairs.
- The structural similarity index [32] is evaluated and stereo pairs with a similarity higher than 70% are discarded, thus leading to a dataset containing 368 pairs.

Cameras, with baseline $b = 5$ mm and focal length $f_c = 863$ px, are calibrated by the `camera_calibration` ROS package which uses the OpenCV calibration function, based on [33]. Subsequently, DMs are created for every pair of RGB images using the `stereo_img_proc` package in which rectification is addressed as detailed in [34]. To validate the calibration process, nine calibrations are evaluated

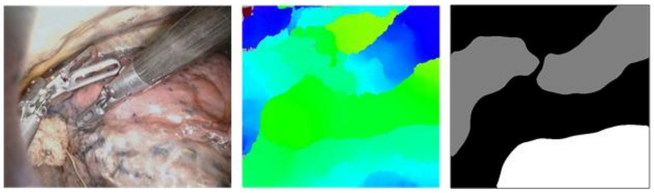

Fig. 3. Example of tool, tissue and background labelling. The coloured DM is manually labelled to highlight the areas containing either a tool (gray) or a candidate tissue flap (white). Note that when the tool touches any anatomical structure, it disappears from the depth map and merges with the background.

and the re-projection error of 0.44 ± 0.06 px is estimated in the projection of the checkerboard points on the image plane. Subsequently, a checkerboard is used to detect four different points showing an error of 7.8 ± 4.4 mm in the 3D estimation.

DMs are manually labelled by means of the MATLAB 2017b Ground Truth Labeler. For a human user, DMs can be difficult to read and understand; therefore, during the labelling process, the user is shown both left and right images in addition to the DM. In every image, two separate labels are created: one representing the tissue flap to be retracted and one representing the DVSS instruments, visible in the scene. Figure 3 shows a sample of endoscopic image (on the left), a DM (in the centre) and a label (on the right). While the purpose of the flap label is to generate the training dataset for the U-Net network, the tool labels are only used to augment the dataset, as described in the following section. The tools' labels are not included in the U-net training set.

C. Dataset Augmentation

The presence of tools in laparoscopic images can obstruct the view and detection of tissue flaps. Moreover, tools introduce a significant disturbance in DMs. In order to enhance the robustness of the U-Net against disturbances generated by tools in the DM, such disturbances must be represented sufficiently in the dataset. An augmentation technique is adopted to improve the network performances. Initially, artificial DMs are generated by extracting the DMs of tools from previously labelled images. Subsequently, portions of the DMs corresponding to the tools are overlapped on images in which no tools were originally present, as shown in Figure 4. With this technique, the dataset initially containing 368 images is increased to 1,080 images. In addition to this technique, random rotation, flipping and zooming are also applied to the dataset using the Keras library [35], thus obtaining a final dataset of 2,160 images.

D. Model Training

The resolution required to identify flaps is lower than the original RGB images produced by the endoscope. Moreover, high resolution images would unnecessarily increase the time required to train the U-Net. Consequently, size of input and target images are reduced from 506x466 (DM valid window) to 64x64, thus allowing for faster training. The network is trained for 200 epochs with a learning rate of 0.001 and

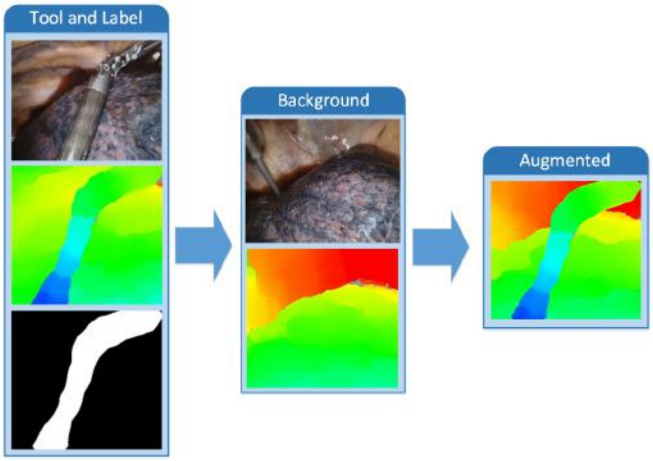

Fig. 4. Augmentation algorithm pipeline. The tool depth map is extracted from the scene (on the left) and superimposed on a depth map where no tools are present or visible (centre). The result is a new image (on the right) which is added to the dataset.

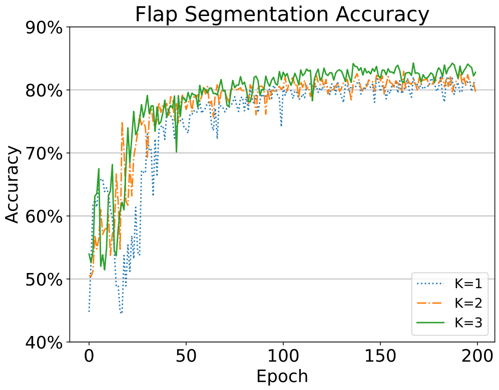

Fig. 5. Accuracy during testing of the K=10 models considered for K-Fold cross-validation. To simplify the data visualisation, only the worst (K=1), the average (K=2) and the best (K=3) cases are shown.

a batch size of 30 images. The Dice loss function [36] is adopted to compute accuracy and the Adam optimiser [37] is used to update the neurons' weights at every epoch.

The augmented dataset is split into a training set (90%) and a test set (10%). In order to assess the robustness of the U-Net against data variability, a training approach based on K-fold [38] cross-validation is adopted. The training process is repeated $K = 10$ times using different subsets of the dataset as training and validation sets.

In Figure 5, the performance of the network over the entire training process is shown for the worst (K=1), average (K=2) and best (K=3) performing model. The network accuracy, defined as the pixel-wise difference between the ground truth and the network prediction, is $80.9\% \pm 1.3\%$ over the K repetitions during the validation phase. The model performance is computed by means of the precision P, defined as $P = \frac{TP}{TP+FP}$ where TP and FP are the true and false positives over the test set respectively. At the end of the training phase, an experimental value of $P = 72.6\% \pm 1.9\%$ is obtained. The network is fed with 64x64 colour depth maps and it outputs 64x64 grayscale masks, with an inference time lower than 42 ms (24 FPS), as measured during the experimental validation phase. The pixel values in the output masks represent the confidence (between 0 and 1) used by the network to identify either the background (0) or the tissue (1). Among the possible detection errors that can affect the U-Net, false positives present the highest risk. In order to reduce the number of false positives, pixels with a confidence value below 80% are classified as background by setting their value to 0 in the tissue mask. The output mask is thus binarised, reducing the noise in the prediction.

E. Gesture execution and planning

After a candidate flap of tissue is identified, the retraction must be planned and subsequently executed. In order to reproduce the gesture, interviews on the standard best practice were conducted with ten experienced clinicians (4 urologists, 3 colorectal surgeons, 2 thoracic surgeons, 1 Ear, Nose and Throat (ENT) surgeon). All clinicians had performed more than 100 robotic surgeries.

Algorithm 1 Retraction planning and execution

```
if tissue not detected then return
else
    $(CT, CB, tissueBorder) = readFromImage();$
    $(sl, inter) = computeLine(CT, CB);$
    $GP = intersection(tissueBorder, sl, inter);$
    $GP = get3DProjection(GP)$
    $(X, Y) = findIntermediatePoint(GP, CB, 25\%)$
    $Z = getQuote(CT) * 1.1$
    $moveTo(X, Y, Z)$
    $align(Z)$
    $OpenGripper()$
    $Z = getQuote(CB)$
    $moveTo(X, Y, Z)$
    while $toolVisible() \lor commandPressed()$ do
        $moveAlong(slope)$
```

From the interviews, the following guidelines emerged:

- The tissue is not grasped; rather,it is mobilised by using the rounded side of the instrument in order to minimise the risk of tissues damage and bleeding.
- The area of interest is the centre of the endoscopic image; therefore, retraction aims to clear the central area from obstructing tissue.
- Instruments approach the surgical area following the direction of the endoscopic view, so to avoid unintentional contact with tissues.
- A suitable point where the instruments approach the tissue is the most central part of the flap, and the retraction is usually performed within the visible area by moving the tissue towards the border of the image.

These guidelines are formalised in the pseudocode reported in Algorithm 1. Based on the labels generated by the U-Net, a set of geometric features is defined (*readFromImage()*). Subsequently, the retraction trajectory is generated (*computeLine(), findIntermedatePoint()*. The cartesian

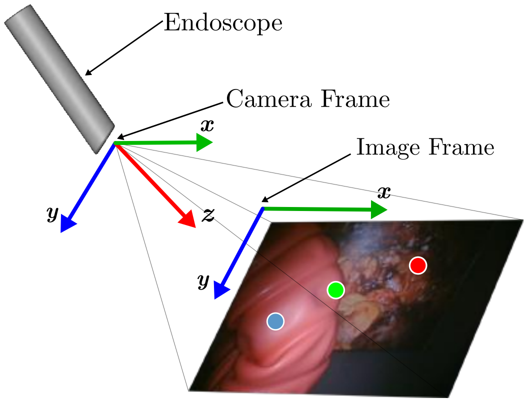

Fig. 6. Representation of the endoscope frame: the X-Y plane of the camera is parallel to the image frame, while the Z axis represent the distance from the origin of the camera frame.

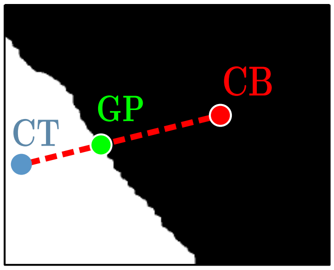

Fig. 7. Feature extracted from the output mask of the U-Net. The tissue (CT) and background point (CB) are estimated as centroids of the areas representing the two classes: tissue flap (white) and background (black). The intersection between the line connecting CT to CB and the edges of the tissue defines the grasping point (GP).

coordinates, shown in Figure 6, are assumed to be in the camera frame - X and Y correspond to the width and height of the image, while the Z coordinate is the depth of the scene based on the direction of the endoscopic view. On the X-Y plane, the centroid of the background (CB, red) and the centroid of the tissue flap (CT, blue) are computed on the b/w image generated by the U-Net, as shown in Figure 7. The grasping point (GP, green) is computed as the intersection between the line connecting the centroids and the border of the tissue (*instersection()*). The 3D position of the aforementioned points is computed by projecting their 2D values on the depth map by applying $Z = \frac{f_c \cdot b}{d}$, where Z is the distance from the camera frame, d is the disparity value of the point, while f_c and b are the camera focal length and baseline respectively (*get3DProjection()*).

Initially, the tool is positioned as follows:

- On the X-Y plane, the tool is positioned in an intermediate position between GP and CB, namely at 25% of the distance in the direction of the GP.
- On Z, the tool pose is set to a z-coordinate evaluated as 0.9 times the distance between $z_C T$ and the camera frame origin, in such a way that it avoids contact with the tissue.
- The tool is aligned along Z.

Subsequently, the tool is moved along Z to the depth of the background and, then, along the direction defined by the line connecting CT to CB. The gesture terminates whether the surgeon releases the pedal or if the tool approaches the boundaries of the image.

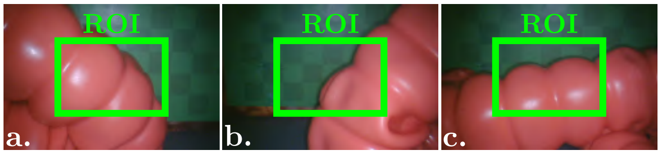

Fig. 8. Examples of initial conditions in the three retraction cases: from the left (a), right (b) and bottom (c). The region of interest (ROI) are highlighted in green.

III. EXPERIMENTAL VALIDATION

A. Experimental Platform

In order to test the approach described above, an experimental platform is used, consisting of the simplified setup shown in Figure 1. A silicone phantom representing a colon is extracted from a training platform for colonoscopy (Kyoto Kagaku M40). A section of the phantom is superimposed on a background image representing the surgical scene, simulating the presence of a tissue flap (i.e. the large bowel) obstructing the surgical view. The network is fed with DMs. Hence, the difference between the surgical images of the training set and the experimental scene has a minimal impact in terms of tissue detection. The platform is placed into a plastic box (36 × 26.5 × 11 cm) to simulate the restricted area available in the abdominal cavity.

Three different scenarios where the bowel segment is placed on the left (Figure 8a), right (Figure 8b) and bottom (Figure 8c) of the scene are investigated to validate the robustness of the flaps detection system as well as the trajectory computation. Every test is repeated 5 times.

The goal of the retraction is to remove tissues obstructing the scene of interest. Hence, a quantitative approach to assess the quality of retraction consists of measuring the area of background image visible after the action is executed. In order to validate the proposed approach, a green checkerboard is superimposed on an endoscopic image of the abdominal cavity, as represented in Fig. 9. The number of visible green background pixels before and after the retraction is evaluated by adopting a Hue Saturation Value (HSV) filter, used as a metric to assess the quality of the procedure. The test is then repeated for a sixth time with the background image without the green checkerboard (i.e. Fig. 9a) to verify that results are comparable. The number of visible pixels is then compared to the number of pixels of an image where no tissue occludes the scene. The same tests are repeated with a background image representing a laparoscopic view, to demonstrate that the algorithm relying on depth maps is affected neither by the presence of the checkerboard, nor by the background.

The hardware setup is composed of a single PSM and a stereo endoscope, as shown in Figure 1. Regarding the computing nodes, a Robot Operating System (ROS)-based network of two computers is used.

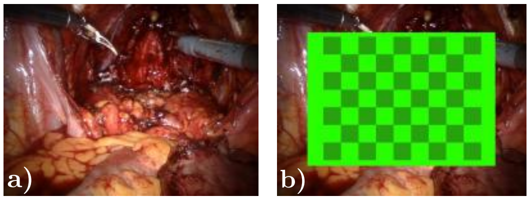

Fig. 9. Different backgrounds used during the tests. The original endoscopic image of abdominal organs (a) and a version with a superimposed green checkerboard (b), used to quantify the amount of background visible before and after the retraction.

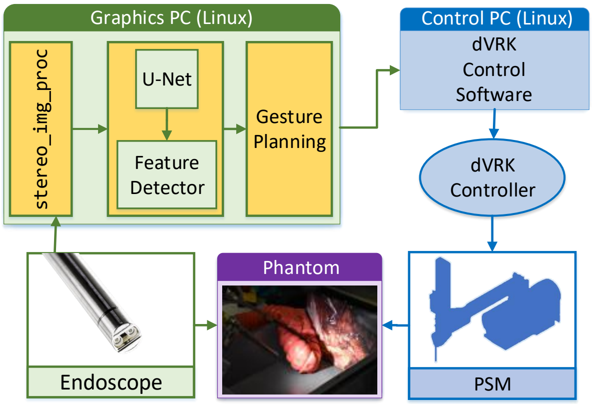

Fig. 10. Experimental setup: DMs are evaluated from endoscopic stereo images and input to the U-Net which estimates the candidate flaps for retraction. Subsequently, features are extracted from the network outputs in order to plan the retraction gesture. Through a DVRK controller installed on a second machine the control is applied to the PSM which performs retraction on the phantom.

The DVRK low level controller, including joint control loops, is installed on a Linux PC (Control PC in Fig. 10) with a ROS interface. This machine is equipped with an Intel Core i5-6400 (2.70 GHz) CPU, HD Graphics 530 and 16GB DD4 (2666 MHz). The computation of the disparity map, the tissue detection U-Net, the feature extraction and the gesture controller are deployed on independent ROS nodes running on a separate machine (Graphics PC in Fig. 10), to prevent instability of the computer running the real-time DVRK controller. The calculator is equipped with an Intel Xeon Gold 6140 (2.30 GHz) CPU, an Nvidia Quadro P1000 GPU, and 128 GB DDR4 2666 MHz RAM. The Da Vinci endoscope used during the tests, calibrated via the procedure detailed in Section II-B, is different from the Da Vinci Xi endoscope used for data collection. The U-Net model used in the detection phase was previously trained on a separate hardware, using the TensorFlow framework.

The surgeon's attention is usually focused on the centre of the surgical scene. For this reason, a region of interest (ROI) is defined as a rectangle placed at the image centre with width and height of half the entire frame. The percentage of visible background is computed for the entire area and for the central ROI.

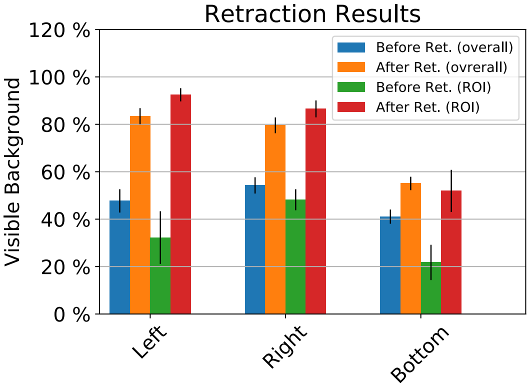

Fig. 11. Field of view enhancement on the entire endoscopic scene expressed in percentage of visible background before and after retraction, accounting for the entire background and the ROI. The performance is calculated as the means over 5 repetition of the three different retraction cases.

B. *Experimental Results*

Numerical results are summarised in Figure 11. Before the retraction, the visible area is $47.7\% \pm 4.9\%$, increasing to $83.4\% \pm 3.3\%$ after the action takes place. On the other hand, the right retraction presents slightly lower performance, increasing from $54.2\% \pm 3.4\%$ to $79.6\% \pm 3.3\%$. This different performance can be explained considering that the PSM is positioned on the left side of the surgical scene, thus performing opposite movements in the two different scenarios. This result suggests that, despite the great dexterity of the DVSS arms, the placement of the PSM with respect to the scene may influence the effectiveness of the retraction.

The worst performance is displayed by the bottom retraction, going from $41.1\% \pm 3.0\%$ to $55\% \pm 2.8\%$. This performance decrease is due to the positioning of the arm, which, similarly to the right retraction case, is subject to a constrained motion. Moreover, the orientation of the arm does not allow the instrument shaft to mobilise the tissue, thus reducing the portion of tool capable of exerting force to the tool tip. The results show that this approach can lead to significant and replicable results. Since the proposed method is new, our results are not comparable with other studies.

The trajectories executed by the DaVinci instrument in the left and right retractions are reported in Figure 12. The solid blue and dashed red lines represent the experiments with and without the checkerboard used as background (Figure 9b and 9a), respectively. The start and end points are shown in green and cyan respectively. The red and blue trajectories are very similar, confirming that background does not significantly affect the task execution. In the trajectories, the different stages of the gesture execution are clearly visible.Although the retraction is planned and executed in separate steps, the last sections of all the trajectories (towards the cyan dot) are very similar and grouped in space, demonstrating that the approach is stable against disturbances and small variations between repetitions. All reported experiments were terminated when the tool reached the edge of the image. Moreover, the different start points (green dots) influence the initial part of the trajectory, before the contact between the tool and tissue takes place. It should be noted that the accuracy of the trajectory execution is completely dependent on the low-level control of the DVRK and is therefore beyond the scope of this work.

IV. CONCLUSIONS

A novel framework for the semi-autonomous planning and execution of tissue retraction is proposed. The combined adoption of deep neural network techniques for image analysis and procedural algorithms for gesture planning is shown as a feasible approach for the execution of autonomous tasks in robotic MIS. Planning and execution of the surgical gesture in the proposed approach can lead to satisfactory and replicable results. The dependability and accuracy of the robot motions offered by this approach can positively impact efficiency, safety and overall user acceptance. Experimental results show an average increase in the visible area of 25% on the whole image and of 42.9% on the ROI. In order to conduct the flap detection stage using a deep learning algorithm, a novel dataset of labelled endoscopic images is developed and released to the community.

To ease the requirement for extensive manual labelling, future developments will concern the adoption of weak labelling [39], unsupervised learning [40] or generative adversarial networks [41] for image segmentation. Improvements in tissue detection may also include procedure-specific detection of organs and the extension of our dataset to images not containing any candidate tissue for retraction. With minor modifications, this will allow to identify when retractable tissue is present in the scene. Detection of large bowel in prostatectomy and liver in cholecystectomy may be beneficial to adjust the parameters of the retraction. Advancements to the procedural algorithm for gesture planning and execution will involve validation on ex-vivo cadaveric models performed by expert surgeons. In addition, further developments to provide a smoother interaction will involve real-time update for the gesture trajectory. The system has been designed in such a way that a clinical DVSS, including the left and right MTMs and two PSMs, can be controlled independently by the surgeon, while the third PSM can be connected to the DVRK control system. As a result, the system can be integrated into a cadaver test for further validation. The system could be combined with a manual laparoscopic procedure or other robotic platforms, where a robotic arm could be used to perform the gesture while the surgeon is operating with conventional instruments.

The main focus of this work is the removal of obstructing tissues in a static scene, which is a simplifying assumption in a realistic scenario. Consequently, future developments should address maintaining the visibility of the surgical area in a dynamic scene and achieving a more accurate depth estimation, possibly by integrating additional sensors and pre-operative analysis. In particular, the on-line evaluation of the visible area is a promising development and will provide an additional step towards its adoption in realistic scenarios. Although the approach described here is developed to reduce interaction with the surgeon, the user interface ergonomics should be considered in the future. A simple yet effective method for displaying the flap and retraction direction is especially required, in conjunction with a robust method for receiving the surgeon's acknowledgement.

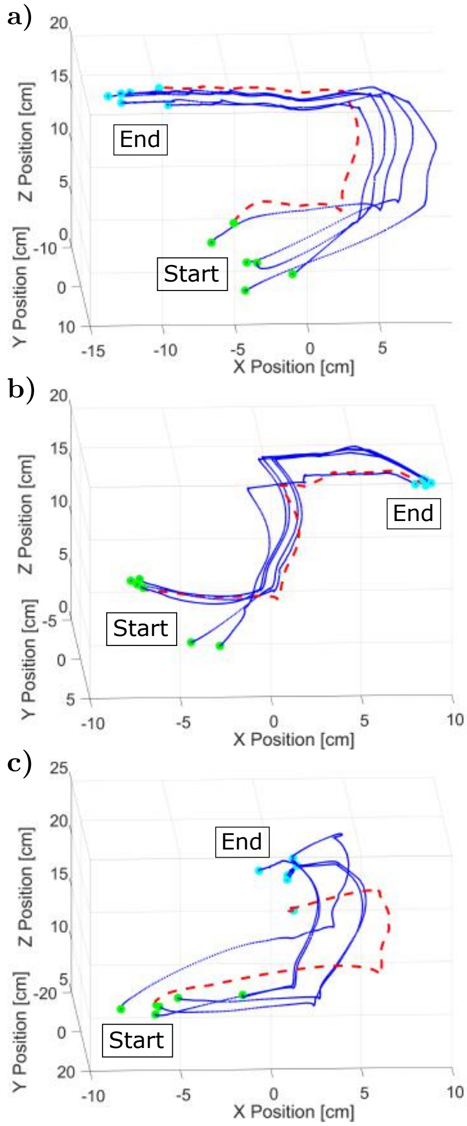

Fig. 12. Tool trajectories during the left (a) right (b) and bottom (c) tissue retraction. The tool starts retracting the phantom tissue from a random position (green). The retraction ends when the tool reaches the edges of the field of view (cyan). Trajectories obtained using the checkerboard background (Fig. 9b) are plotted in blue, while the control experiments performed with the endoscopic background (Fig. 9a) are plotted with a dashed red line.

ACKNOWLEDGEMENTS

Research reported in this article was supported by Intuitive Surgical Inc. under the Technology Research Grants program 2019, by the Royal Society, by the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R045291/1, and by the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 818045).

Any opinions, findings and conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Royal Society, EPSRC, or the ERC. The authors thank Intuitive Surgical, and in particular Simon Di Maio and Dale Bergman, for supportng the research and allowing us to work on the dVRK. Alejandro F. Frangi acknowledges support from the Royal Academy of Engineering Chair in Emerging Technologies Scheme (CiET1819/19).

References

[1] Z. Moghadamyeghaneh, M. H. Hanna, J. C. Carmichael, A. Pigazzi, M. J. Stamos, and S. Mills, "Comparison of open, laparoscopic, and robotic approaches for total abdominal colectomy," *Surgical Endoscopy*, vol. 30, no. 7, pp. 2792–2798, jul 2016.

[2] P. Steele *et al.*, "Current and future practices in surgical retraction," *The Surgeon*, vol. 11, no. 6, pp. 330–337, dec 2013.

[3] M. Liu and M. Curet, "A Review of Training Research and Virtual Reality Simulators for the da Vinci Surgical System," *Teaching and Learning in Medicine*, vol. 27, no. 1, pp. 12–26, 2015.

[4] K. Catchpole *et al.*, "Safety, efficiency and learning curves in robotic surgery: a human factors analysis," *Surgical Endoscopy*, vol. 30, no. 9, pp. 3749–3761, sep 2016.

[5] J. C. Hu *et al.*, "Perioperative complications of laparoscopic and robotic assisted laparoscopic radical prostatectomy," *The Journal of urology*, vol. 175, no. 2, pp. 541–546, 2006.

[6] 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*. Cham: Springer International Publishing, 2015, pp. 234–241.

[7] M. Benkhadra *et al.*, "Flexibility of thiel's embalmed cadavers: the explanation is probably in the muscles," *Surgical and Radiologic Anatomy*, vol. 33, no. 4, pp. 365–368, May 2011.

[8] 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 *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2014, pp. 6434–6439.

[9] B. S. Peters, P. R. Armijo, C. Krause, S. A. Choudhury, and D. Oleynikov, "Review of emerging surgical robotic technology," *Surgical Endoscopy*, vol. 32, no. 4, pp. 1636–1655, Apr 2018.

[10] S. Bodenstedt *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," *arXiv:1805.02475*, 2018.

[11] A. Krieger *et al.*, "Development and Feasibility of a Robotic Laparoscopic Clipping Tool for Wound Closure and Anastomosis," *Journal of Medical Devices*, vol. 12, no. 1, 2017.

[12] S. McKinley *et al.*, "An interchangeable surgical instrument system with application to supervised automation of multilateral tumor resection," *IEEE International Conference on Automation Science and Engineering*, vol. nov 2016, pp. 821–826, 2016.

[13] Muradore *et al.*, "Development of a cognitive robotic system for simple surgical tasks," *International Journal of Advanced Robotic Systems*, vol. 12, 2015.

[14] H. Nakawala, R. Bianchi, L. E. Pescatori, O. De Cobelli, G. Ferrigno, and E. De Momi, "Deep-Onto network for surgical workflow and context recognition," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 4, pp. 685–696, 2019.

[15] B. Kehoe *et al.*, "Autonomous multilateral debridement with the Raven surgical robot," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2014, pp. 1432–1439.

[16] S. Sen *et al.*, "Automating multi-throw multilateral surgical suturing with a mechanical needle guide and sequential convex optimization," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2016, pp. 4178–4185.

[17] C. D'Ettorre *et al.*, "Automated pick-up of suturing needles for robotic surgical assistance," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 1, no. c, pp. 1370–1377, 2018.

[18] Y. Gao *et al.*, "JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): A Surgical Activity Dataset for Human Motion Modeling," *Modeling and Monitoring of Computer Assisted Interventions*, pp. 1–10, 2014.

[19] A. Murali *et al.*, "Learning by observation for surgical subtasks: Multilateral cutting of 3D viscoelastic and 2D Orthotropic Tissue Phantoms," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015-June, no. June, pp. 1202–1209, 2015.

[20] R. Jansen, K. Hauser, N. Chentanez, F. Van Der Stappen, and K. Goldberg, "Surgical retraction of non-uniform deformable layers of tissue: 2D robot grasping and path planning," *IEEE International Conference on Intelligent Robots and Systems, IROS 2009*, pp. 4092–4097.

[21] S. Patil and R. Alterovitz, "Toward automated tissue retraction in robot-assisted surgery," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2088–2094, 2010.

[22] R. Elek *et al.*, "Towards surgical subtask automation-blunt dissection," in *INES 2017 - IEEE 21st International Conference on Intelligent Engineering Systems, Proceedings*, 2017, pp. 253–257.

[23] T. D. Nagy, M. Takacs, I. J. Rudas, and T. Haidegger, "Surgical subtask automation - Soft tissue retraction," in *2018 IEEE 16th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE, feb 2018, pp. 55–6.

[24] F. Isensee *et al.*, "Brain Tumor Segmentation Using Large Receptive Field Deep Convolutional Neural Networks," in *Brain Tumor Segmentation Using Large Receptive Field Deep Convolutional Neural Networks*, 2017, pp. 86–91.

[25] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation From CT Volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, dec 2018.

[26] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep Learning Based Robotic Tool Detection and Articulation Estimation with Spatio-Temporal Layers," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2714–2721, 2019.

[27] Fedorov *et al.*, "3D Slicer as an image computing platform for the Quantitative Imaging Network," *Magnetic Resonance Imaging*, vol. 30, no. 9, pp. 1323–1341, nov 2012.

[28] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org.

[29] J. Ko, M. Kim, and C. Kim, "2d-to-3d stereoscopic conversion: depth-map estimation in a 2d single-view image," in *Proc. SPIE 6696, Applications of Digital Image Processing XXX*, sep 2007.

[30] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 328–341, 2007.

[31] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[32] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[33] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, 2000.

[34] R. I. Hartley, "Theory and practice of projective rectification," *International Journal of Computer Vision*, vol. 35, no. 2, pp. 115–127, 1999.

[35] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[36] F. Milletari *et al.*, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision (3DV)*, 2016, pp. 565–571.

[37] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proceedings of the 3rd International Conference on Learning representations*, dec 2015, pp. 1–15.

[38] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[39] F. Fuentes-Hurtado, A. Kadkhodamohammadi, E. Flouty, S. Barbarisi, I. Luengo, and D. Stoyanov, "EasyLabels: weak labels for scene segmentation in laparoscopic videos," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1247–1257, 2019.

[40] T. Moriya *et al.*, "Unsupervised segmentation of 3D medical images based on clustering and deep representation learning," in *Medical Imaging*, 2018, p. 71.

[41] A. Rau *et al.*, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1167–1176, 2019.