Evaluation of Methods for Semantic Segmentation of Endoscopic Images

1st Bastian Bopp
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Karlsruhe, Germany
uedxg@student.kit.edu

2nd Paul Maria Scheikl
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Karlsruhe, Germany
paul.scheikl@kit.edu

3rd Christian Kunz
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Karlsruhe, Germany
christian.kunz@kit.edu

4th Franziska Mathis-Ullrich
Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology
Karlsruhe, Germany
franziska.ullrich@kit.edu

***Abstract*—We examined multiple semantic segmentation methods, which consider the information contained in endoscopic images at different levels of abstraction in order to predict semantic segmentation masks. These segmentations can be used to obtain position information of surgical instruments in endoscopic images, which is the foundation for many computer assisted systems, such as automatic instrument tracking systems. The methods in this paper were examined and compared in regard to their accuracy, effort to create the data set, and inference time. Of all the investigated approaches, the LinkNet34 encoder-decoder network scored best, achieving an Intersection over Union score of 0.838 with an inference time of 30.25 ms on a 640 x 480 pixel input image with a NVIDIA GTX 1070Ti GPU.**

***Index Terms*—robot-assisted surgery, computer vision, semantic image segmentation, deep learning**

I. Introduction

During minimally invasive surgery the surgeon often has no direct view of the surgical site within the human body but utilises an endoscope which is guided by a human assistant. The optimal positioning of that endoscope is a complex problem, which in practice often leads to poor visual conditions and interruptions of the surgical workflow. However, high-quality camera images are crucial for surgical performance [1]. This surgical challenge results in the need for an automatic tracking system, which recognises the presence and position of surgical instruments in visual data and is able to control the endoscope's position based on this information [2].

Early approaches to classical image processing, which use color and shape information from tools or attached markers, are not robust against common challenges in endoscopic image data, such as fluctuating exposure conditions, light reflections, occlusions, smoke generated by electrocautery, repetitive textures with low contrast, and large deformations of anatomical structures. Machine learning approaches are often more robust than classical image processing methods in regard to these challenges.

In this work, a variety of machine learning methods were examined in regard to their suitability for solving these problems in the task of creating binary segmentation masks to classify image pixels as belonging to instrument or background. The investigated methods differ significantly in complexity, the type of features used to solve the task, and the amount of data required for training. Gathering data in the medical field is expensive and time consuming. Especially for pixel-wise segmented endoscopic images.

II. Methods

The following sections describe briefly the investigated approaches for creating binary segmentation masks (instrument/background). For this task, weakly supervised (II-A) and fully supervised (II-B, II-C, II-D) approaches were examined in order to compare their performance. The data set used to train and validate the following methods is comprised of data from [3] and [4] and consists of 420 pixel-wise annotated endoscopic images.

A. Class Activation Mapping

Class Activation Mapping is a method which can be used to visualise the predicted class scores of a classification-trained convolutional neural network (CNN) on any given image [5]. Those image regions can be used to perform object localisation and are obtained in a weakly-supervised manner using classical CNNs trained on image-level labels (weak-labels).

B. Random Forest Segmentation

A Random Forest (RF) is trained as a binary (instrument/background) classifier for image pixels. The features used for the segmentation of endoscopic instruments are the pixel values in hue, saturation, opponent 2 and opponent 3 color spaces, which are superior to other color spaces like RGB or CIE XYZ concerning discriminative power between instruments and background [6].

C. Fully Convolutional Network

A shallow Fully Convolutional Network (FCN) (3 layers deep and 6331 trainable parameters) is trained to produce binary segmentation masks. The investigated architecture uses rectangular image regions of 7×7 pixels to predict the class of one pixel centered in the image region. Hereby, in contrast to the RF, the model is more robust against pixel noise in the image data.

D. Encoder-Decoder Network

Our final method is a state-of-the-art encoder-decoder network that we trained on solving the problem of semantic segmentation. The architecture used in this work is based on the LinkNet model [7] and utilises a pre-trained ResNet34 [8] network as an encoder similar to [9].

III. Results

A. Weakly Supervised Segmentation

Fig. 1 shows exemplary endoscopic images overlayed with their corresponding class activation maps (CAMs). Computing a CAM for a 224×224 pixel input image takes 3500 ms on a single NVIDIA GTX 1070Ti GPU.

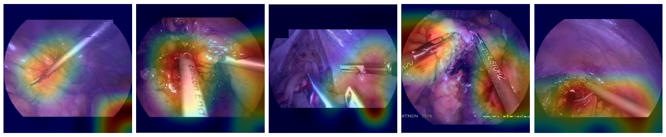

Fig. 1. Exemplary class activation maps.

The examples in Fig. 1 illustrate that classical CNNs, that were trained on weakly labelled endoscopic image data, are able to not only detect the presence of instruments but also localise the image regions in which they appear. However, the sole use of CAMs results in relatively coarse detected image regions and fails to detect instrument boundaries. Therefore this method is ineligible to create semantic segmentation masks for endoscopic images on its own.

B. Fully Supervised Segmentation

The qualitative comparison of the investigated fully supervised segmentation approaches is presented in Table I and Fig. 2.

TABLE I
Binary segmentation results and inference time.

Model	Intersection over Union	Dice coefficient	Time [ms]
RF	0.431	0.602	841.90
FCN	0.529	0.692	**7.50**
LinkNet34	**0.838**	**0.912**	30.25

The inference times of the FCN and LinkNet34 models were measured using a single NVIDIA GTX 1070Ti GPU. The current implementation of the RF segmentation model does not allow GPU processing, which results in a significantly higher inference time. All metrics were obtained on 640×480 pixel images. Fig. 2 shows examples of typical endoscopic images and their corresponding predicted segmentation masks.

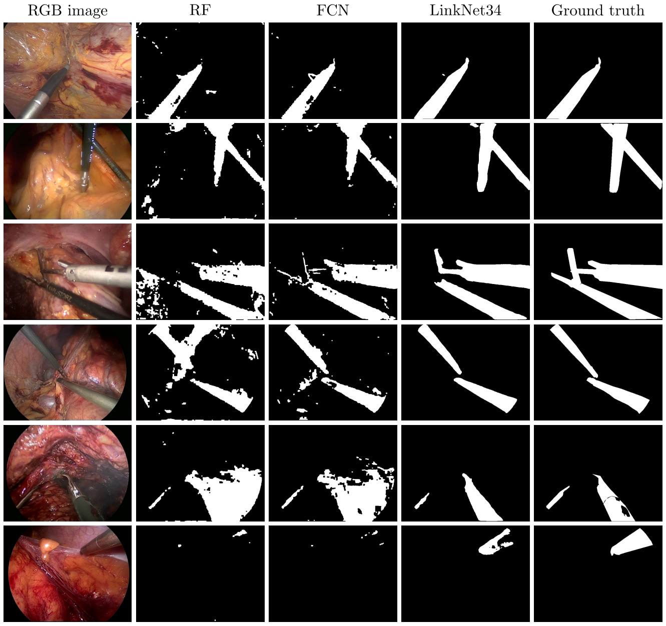

Fig. 2. Qualitative comparison between the predicted binary segmentation masks of different models.

Table I and Fig. 2 show that the deep LinkNet34 encoder-decoder architecture, which considers the whole image at once to predict its segmentation masks, considerably outperforms the smaller models that have much fewer trainable parameters and segment an image based on the pixel values of individual pixels (RF) or small image regions (FCN).

IV. Conclusion

In principle, all of the investigated methods are able to recognise the position of the instruments occurring in endoscopic image data qualitatively. The weakly supervised approach of Class Activation Mapping is appealing, since no pixel-wise segmented training data is required. However, the approach lacks segmentation accuracy to be used as a stand-alone method for semantic segmentation of surgical instruments. The relatively small RF and FCN models require little training data and computing time (with appropriate implementation), but they are not robust against challenges such as smoke or contaminated instruments. Herein lies the strength of the deep LinkNet34 encoder-decoder network, which however requires large amounts of annotated data for training.

In future work, class activation maps will be used as seed points for additional segmentation methods. Furthermore, the combination of active learning methods with segmentation networks will be investigated in order to further increase the accuracy of the methods and to limit the annotation effort at the same time.

References

[1] R. D. Shah, et al., “Performance of basic manipulation and intracorporeal suturing tasks in a robotic surgical system: single- versus dual-monitor views,” Surgical Endoscopy, vol. 23, no. 4, pp. 727–733, Apr 2009.

[2] X. Du, et al. “Articulated multi-instrument 2-D pose estimation using fully convolutional networks,” IEEE transactions on medical imaging 37.5, 2018.

[3] L. Maier-Hein, et al., “Can masses of non-experts train highly accurate image classifiers?” in Medical Image Computing and Computer-Assisted Intervention MICCAI 2014, P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, Eds. Cham: Springer International Publishing, 2014, pp. 438-445.

[4] MICCAI 2017 Endoscopic Vision Challenge, “Robotic Instrument Segmentation Sub-Challenge,” https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization.” CVPR, 2016.

[6] M. Allan, et al., “Toward detection and localization of instruments in minimally invasive surgery,” IEEE Transactions on Biomedical Engineering, vol. 60, no. 4, pp. 1050–1058, April 2013.

[7] A. Chaurasia and E. Culurciello, “Linknet: Exploiting encoder representations for efficient semantic segmentation,” CoRR, vol. abs/1707.03718, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CoRR, vol. abs/1512.03385, 2015.

[9] A. Shvets, A. Rakhlin, A. A. Kalinin, and V. Iglovikov, “Automatic instrument segmentation in robot-assisted surgery using deep learning,” CoRR, vol. abs/1803.01207, 2018.