A Survey of Semantic Segmentation

Marvin Teichmann* Martin Thoma*

***Abstract*—Distinguishing medical instruments from background is a basic task which supplements many applications and further analysis. If this task works robust and fast, one can think about instance segmentation of medical instruments, operation phase detection, pose estimation of the instruments, recognition of the organs and ill tissue.**

This work is part of a practical at KIT.

I. Introduction

Operations like the removal of a tumor or the gallbladder (a cholecystectomy) require the body of the patient to be opened. However, minimal-invasive operations got more and more attentions since 1987 [Wic87]. In this kind of operation, the surgeon tries to make as little and as small cuts in the patients body as possible. The advantage is that the patients skin can heal faster and thus the patient can recover faster from the damage which was done by the operation. The disadvantage of minimal-invasive operations is that the operation itself gets harder for the surgeon. Special medical equipment has to be used: Small cameras and fiber optic cables so that the surgeons can see what their doing,

Machines can not only provide the possibility to make more fine-grained movements (e.g. with the *da Vinci* Surgical System (Intuitive Surgical, Mountain View, Calif)), but also improve vision. For example, the limited visibility due to cautarization-induced smoke can be fought by highlighting the medical instruments, the instruments themselves can be recognized and the operation phase can be detected. If the camera images had a pixel-wise segmentation of medical instruments and background, those tasks would be simpler.

II. Related Work

Pixel-wise segmentation was successfully applied in other domains. For example, [BKTT15] introduces a system which does pixel-wise segmentation for autonomous cars of the classes *street* and *no street*. A more detailed introduction to semantic segmentation can be found in [Tho16].

III. Experiments

The experiments were done on the *Instrument segmentation and tracking* dataset of the Endoscopic Vision Challenge "EndoVis"[1]. It contains photos of minimal-invasive operations. The dataset already contains a training-test-split. The training data consists of four operations. Each operation has 40 RGB images in a resolution of $640\,\text{px} \times 480\,\text{px}$. The test set has 10 more images of those 4 operations as well as 50 images for 2 more operations. This means, in total the training data consist of $4 \cdot 40 = 160$ photos and the testing contains $4 \cdot 10 + 2 \cdot 50 = 140$ photos. As each pixel has to be classified as *medical instrument* or *background*, there are $140 \cdot 640 \cdot 480 = 43\,008\,000$ classifications to be done for testing.

A desktop computer with a Titan Black and an Intel Core i7-4930K was used for evaluation.

A. Base line experiments

TODO: Constant background? Accuracy, Precision, Recall?

The most basic information for pixel-wise semantic segmentation is the color of the pixel. Typically, images are in RGB format. This means the image has three channels (Red, Green, Blue). Each channel has 8 bit and thus $2^8 = 256$ possible values, ranging from 0 to 255. This gives $(2^8)^3 = 16\,777\,216$ possible colors. Obviously, only the color can not give a perfect result in all circumstances as the measured color changes due to smoke, shaddows, specular highlights and insufficient illumination. But it gives an impression how important local features are for the specific problem.

A model with 64 sigmoid nodes in a first hidden layer with $50\,\%$ dropout, 64 sigmoid nodes in a second hidden layer with $50\,\%$ and one sigmoid output unit achieved a pixel-wise accuracy of $92.88\,\%$[2], a precision of $76.13\,\%$ and a recall of $32.94\,\%$. The confusion matrix is given in Table I.

B. Local Models

A slightly better model than the baseline used the coordinate of the pixel in the image as well as $3\,\text{px}$ erosion and dilation. This is known as a morphological opening. This model achieved an per-pixel accuracy of $93.51\,\%$, a precision of $76.74\,\%$ and a recall of $42.30\,\%$. The confusion matrix is given in Table III.

If only the opening operation was applied, then the model achieves only an accuracy of $91.07\,\%$, a precision of $74.76\,\%$ and a recall of $4.50\,\%$.
If the model with coordinates is used without morphological opening, then it gets an accuracy of $93.42\,\%$, a precision of $76.98\,\%$ and a recall of $40.65\,\%$.

* These authors contributed equally to this work.

[1] http://endovis.grand-challenge.org

[2] This is the same as the DICE coefficient.

Astonishingly, the accuracy of models which made use of a tiny patch around the pixel which was to classify decreased to about $92.11\,\%$, the precision is $42.30\,\%$ and the recall is $76.74\,\%$. The classification of one image took about $1.54\,\mathrm{s}$.

TODO: This may not happen. The model could simply set the weights to 0 and thus ignoring the other features. Where exactly is the problem?

C. Fully Convolutional Networks

Fully Convolutional Networks (FCNs) were introduced by [LSD15]. They were shown to improve performance in semantic segmentation tasks over standard convolutional neural networks (CNNs).

TODO

IV. Discussion

TODO

V. Acknowledgement

We would like to thank the "Begabtenstiftung Informatik Karlsruhe" for supporting our research.

References

[BKTT15] S. Bittel, V. Kaiser, M. Teichmann, and M. Thoma, "Pixel-wise segmentation of street with neural networks," *arXiv preprint arXiv:1511.00513*, 2015. [Online]. Available: http://arxiv.org/abs/1511.00513

[LSD15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440. [Online]. Available: http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=7478072

[Tho16] M. Thoma, "A survey of semantic segmentation," *arXiv preprint arXiv:1602.06541*, Feb. 2016. [Online]. Available: http://arxiv.org/abs/1602.06541

[Wic87] J. E. Wickham, "The new surgery." *British medical journal (Clinical research ed.)*, vol. 295, no. 6613, p. 1581, Dec. 1987.

APPENDIX A
TABLES

		Predicted class	
		0	**1**
Actual	**0**	38 640 407	2 654 875
class	**1**	408 816	1 303 902

Table I: Confusion matrix of a model trained soley on pixel colors. Class 0 is background, class 1 is medical instruments.

		Predicted class	
		0	**1**
Actual	**0**	38 989 042	3 780 484
class	**1**	60 181	178 293

Table II: Confusion matrix of a model trained soley on pixel colors. Additionally, a morphological closing operation with $3\,\mathrm{px}$ was applied. Class 0 is background, class 1 is medical instruments.

		Predicted class	
		0	**1**
Actual	**0**	38 541 570	2 284 099
class	**1**	507 653	1 674 678

Table III: Confusion matrix of a model trained on pixel colors and the pixels coordinates. Additionally, a morphological closing operation was applied. Class 0 is background, class 1 is medical instruments.