Image-based segmentation and localization of surgical instruments using deep neural networks

Amelie Wagner

***Abstract*—Segmentation and localization of surgical instruments in endoscopic videos during minimally invasive surgery is a current challenge for evaluating and supporting the performance of the surgeon.**

For the instrument segmentation task, every pixel of the input image has to be labeled as background or instrument. For the instrument localization task, the position of the instrument center point in the image has to be estimated. For the instrument segmentation task, every pixel of the network input image is labeled as black for background or white for instrument. In this work, both tasks are solved jointly by an adjusted TernausNet-11 network. For the localization task, the TernausNet-11 predicts a heatmap centered at the position of the instrument center point. Afterwards, the center point coordinates of each instrument are extracted by using weighted k-means clustering. Every model proposed in this work is evaluated on the robotic dataset of the MICCAI 2015 Endoscopic Vision, on the subchallenge instrument segmentation and tracking according to the challenge guidelines. The evaluation results show that the proposed method is able to solve both tasks on the same level as state-of-the-art methods.

I. Introduction

Minimally invasive surgery (MIS) is a surgical technique that has several advantages in comparison to the commonly used standard open approach, for example shorter recovery time and hospital stays after the intervention [1]. Because the field of view for the surgeon is reduced by the endoscopic camera and the instruments can only be seen two-dimensional, it is important to support the surgeon during this procedure [2]. Segmenting and locating the instruments directly out of the endoscopic video images is an attractive possibility for this, especially because no modification of the surgical scene is required. The goal of this work is the segmentation and localization of surgical instruments out of endoscopic video images. For the segmentation task, the image is partitioned into black pixels for background and white pixels for instrument. For the localization task, the coordinates of the instrument center point are extracted out of the image. In this work, a Convolutional Neural Network (CNN) is used for segmenting and localizing the instrument. It is based on TernausNet-11 [3] that already solves the segmentation task. The resulting localization network is able to solve the localization task based on the information learned for segmentation.

II. Methods

A. Instrument Localization Network

The segmentation network TernausNet-11 is extended by three more layers to solve the localization task. It predicts two outputs, one for the segmentation and one for the localization task. The output of the segmentation output layer is a binary image, that contains a white segmentation mask where the instrument is located. The output for the localization task is a greyscale heatmap. A high pixel value at the predicted heatmap indicates a high probability that the instrument center point is located at this pixel position. The network is abbreviated as *LocNet*.

B. Heatmaps

It is necessary to convert the localization targets, given as two-dimensional image positions, into greyscale images, in order to make it possible for LocNet to process them. The heatmaps are generated by calculating a two-dimensional Gaussian distribution $gauss2D(x_r, y_r)$, centered at the center point (x_{cp}, y_{cp}) of the instrument.

$$H(x_r, y_r) = gauss2D(x_r, y_r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_r - x_{cp})^2 + (y_r - y_{cp})^2}{2\sigma^2}} \tag{1}$$

The standard deviation σ controls the spread of the Gaussian around the instrument location. σ^2 denotes the variance. By iterating over the dimensions of the training input image and applying $gauss2D(x_r, y_r)$, a heatmap with the same dimensions is generated out of the ground truth instrument position. (x_r, y_r) is one pixel position of the training image, which corresponds to the pixel position in the generated heatmap. When two instruments are visible in the input image, the heatmap is generated by calculating two Gaussian distributions, one for each instrument

III. Postprocessing

The heatmaps predicted by the network mostly have a pixel range from $[127, 255]$. In order to improve the extraction of the instrument position, the pixel values of the predicted heatmaps are thresholded with $threshold$ set to 129:

$$thresh(H(x_r, y_r)) = \begin{cases} H(x_r, y_r) & \text{if } H(x_r, y_r) \geq threshold \\ 0 & \text{otherwise} \end{cases}$$

In a second postprocessing step, histogram equalization [4] is used to increase the contrast of the Gaussian distribution in the image.

To get the instrument position out of the thresholded heatmap, *k-means* clustering is used. It is a commonly used method to automatically partition a dataset into k groups [5]. When the input to the corresponding predicted heatmap contains one instrument, k is set to 1. When two instruments are visible in the input image, k is set to 2. The pixel values of

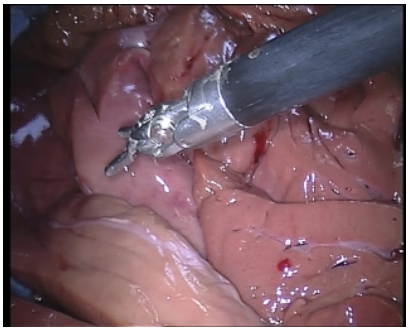

Fig. 1. Example image out of Endoscopic Vision Challenge 2015.

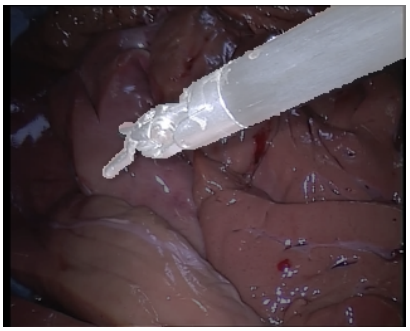

Fig. 2. Example image out of Endoscopic Vision Challenge 2015 with superimposed segmentation mask.

the thresholded heatmap are used as weights for the k-means clustering to improve the results, because a higher pixel value indicates a higher probability that the instrument center point is located at that position.

IV. Evaluation

A. Endoscopic Vision Challenge 2015

The datasets used for segmentation and localization are proposed in the subchallenge Instrument Segmentation and Tracking of the MICCAI Endoscopic Vision Challenge 2015 [6]. The subchallenge is abbreviated as *EndoVis15*. The robotic set of the subchallenge that is used in this work, is separated in a segmentation and a localization part. One input image has two different ground truth annotations, one for each subchallenge part. The segmentation part is abbreviated as *EndoVis15-S*, the tracking part is abbreviated as *EndoVis15-T*.

B. Experiments

After preprocessing the proposed datasets, LocNet was trained and evaluated in different ways with the EndoVis-15 datasets. Each model is trained and evaluated according to one of the EndoVis15 challenge guidelines: On each of the four surgeries, one model is trained. Then it is tested with the remaining three surgeries (LOSO). The other evaluation method consists of training on all train sets and then testing on the test sets (T4). The results of the best models for each evaluation method are proposed in table I and table II. The models were trained to the conditions stated in table III.

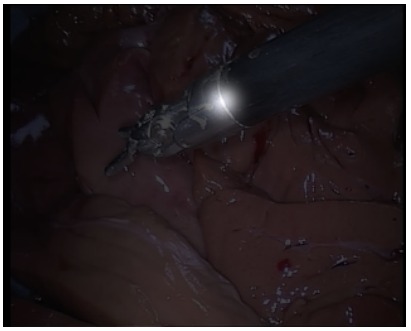

Fig. 3. Example image out of Endoscopic Vision Challenge 2015 with superimposed heatmap.

Best Localization Results LOSO				
Dataset1 left/right instr.	Dataset2	Dataset3	Dataset4	mean dist.
17.58/15.15	11.36	10.83	12.05	13.85

TABLE I

Results for the different datasets for the best LOSO LocNet model. Mean dist. is the mean distance over all datasets. Dataset1 is distinguished into left and right instrument (left/right instr.). The evaluation results are given as mean value over all prediction results. The distance between ground truth location and predicted location of the instrument is given in pixels.

V. Conclusion

The achieved results are comparable to other state-of-the-art approaches. Besides taking further postprocessing steps to improve results, the proposed method could be combined with a temporal tracking algorithm. In this case, the estimated instrument locations could serve as input to trackers such as Kalman filter or Particle filter [7]. Temporal tracking would be especially helpful to deal with occlusions and to associate each heatmap with either the left or the right instrument, even when instruments cross.

Localization Results Datasets T4		
Dataset5 left/right instr.	Dataset6 left/right instr.	mean dist.
89.80/117.91	88.17/120.02	106.4

TABLE II

Results of the best T4 LocNet model. Each dataset is distinguished into left and right instrument (left/right instr.). Mean dist. is the mean distance over all datasets. The evaluation results are given as mean value over all prediction results $\pm$ the standard deviation. The distance between ground truth location and predicted location of the instrument is given in pixels.

Training Conditions LocNet		
val. method	epochs	γ
LOSO model	50	1/2
T4 model	500	1/2

TABLE III

Validation method (val.method) specifies if the model was trained according to the LOSO fashion (LOSO), or trained completely on the four training sets and tested on the two test sets (T4). The impact of the segmentation and the localization loss on the model is adjusted by γ.

REFERENCES

[1] A. M. Lacy, J. C. Garca-Valdecasas, S. Delgado, A. Castells, P. Taur, J. M. Piqu, and J. Visa, "Laparoscopy-assisted colectomy versus open colectomy for treatment of non-metastatic colon cancer: a randomised trial," *The Lancet*, pp. 2224 – 2229, 2002.

[2] M. Allan, S. Ourselin, S. Thompson, D. J. Hawkes, J. Kelly, and D. Stoyanov, "Toward Detection and Localization of Instruments in Minimally Invasive Surgery," *IEEE Transactions on Biomedical Engineering*, pp. 1050–1058, 2013.

[3] A. Shvets, A. Rakhlin, A. A. Kalinin, and V. Iglovikov, "Automatic Instrument Segmentation in Robot-Assisted Surgery Using Deep Learning," 2018.

[4] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, pp. 355 – 368, 1987.

[5] D. A. Sipkins, X. Wei, J. W. Wu, J. M. Runnels, D. Côté, T. K. Means, A. D. Luster, D. T. Scadden, and C. P. Lin, "In vivo imaging of specialized bone marrow endothelial microdomains for tumour engraftment," *Nature*, p. 969, 2005.

[6] "MICCAI Endoscopic Vision Challenge: Subchallenge Instrument Segmentation and Tracking," https://endovissub-instrument.grand-challenge.org/, accessed: 2018-09-15.

[7] R. G. Brown, P. Y. Hwang *et al.*, *Introduction to random signals and applied Kalman filtering*. Wiley New York, 1992.