Autonomous Neurosurgical Instrument Segmentation Using End-to-End Learning

Niveditha Kalavakonda, Blake Hannaford
University of Washington
Seattle, WA
nkalavak, blake @uw.edu

Zeeshan Qazi, Laligam Sekhar
Harborview Medical Center
Seattle, WA
zeeshan, lsekhar@neurosurgery.washington.edu

Abstract

Monitoring surgical instruments is an essential task in computer-assisted interventions and surgical robotics. It is also important for navigation, data analysis, skill assessment and surgical workflow analysis in conventional surgery. However, there are no standard datasets and benchmarks for tool identification in neurosurgery. To this end, we are releasing a novel neurosurgical instrument segmentation dataset called NeuroID for advancing research in the field. Delineating surgical tools from the background requires accurate pixel-wise instrument segmentation. In this paper, we present a comparison between three encoder-decoder approaches to binary segmentation of neurosurgical instruments, where we classify each pixel in the image to be either tool or background. A baseline performance was obtained by using heuristics to combine extracted features. We also extend the analysis to a publicly available robotic instrument segmentation dataset and include its results. The source code for our methods and the neurosurgical instrument dataset will be made publicly available[1] *to facilitate reproducibility.*

1. Introduction

Automated instrument identification from surgical video frames will aid the development of intelligent applications in instrument tracking, pose estimation, augmented reality overlay, visual servoing, analysis of surgical phase, and will aid development of safety mechanisms for surgery. In addition to benefiting different components in surgery, the position information of instruments also helps in measuring the context-awareness of a surgeon, potentially helping to reduce human errors. Bounding box recognition of instruments is not sufficient due to the coarse boundaries generated around the regions of interest. The dense prediction of tool versus background through semantic segmentation enhances safety by parsing the whole scene without suppressing occluded or slanted objects.

Segmentation of instruments is a challenging task due to variation in lighting, shadows, instrument overlap, reflections from tissue on instrument (and vice versa), a range of textures, and occlusions from blood, fogging or surgeon interference in the frame[3]. The diverse range of surgical instruments, resolution of images/videos captured and conditions in the surgical space (occlusions, rapid appearance changes, specular reflections and blur) also affect the robustness of tool detection and identification.

Related Work: [7] uses template-matching to compensate the offset error from robot kinematics. Jo et. al[6] make use of the L*A*B color space, histogram equalization and Otsus thresholding to segment instruments in robot-assisted laparoscopic surgery images. Methods like Maximum likelihood Gaussian Mixture Models and Naive Bayesian classifiers (summarized in [3]) can be used to classify pixels of surgical tools and background. However, these methods require considerable processing in advance to determine features. Bouget et. al[4] detect surgical tools by learning the local appearance and global shape from the training data. They trained a random forest model over ten feature channels and the shape model is learned using a linear Support Vector Machine. The parameters were searched exhaustively making it less useful for run-time implementation. Recent developments in deep-learning have showed dramatic improvements in segmenting instruments with increased accuracy and real-time performance[2].

Contribution: We generated a labeled dataset for neurosurgical instrument segmentation and identification (NeuroID, short for Neurosurgical Instrument Dataset). The dataset provides pixel-wise annotation for tool versus background and includes the class of each instrument. The images were manually annotated using pre-determined labels for the classes. We have also incorporated variance in conditions by choosing frames from across different surgical procedures. To evaluate initial performance on the dataset,

[1] http://brl.ee.washington.edu/robotics/surgical-robotics/neurosurgical-instrument-segmentation/

we report comparison of four different automated methods for binary segmentation of neurosurgical instruments. Some of our approaches were also evaluated on a public robotic instrument dataset[1] for comparison.

2. Datasets

Neurosurgical Instrument Dataset: The NeuroID dataset has been generated from five videos recorded *in vivo* at the Harborview Medical Center in Downtown Seattle. An application to the Institutional Review Board at the University of Washington was approved (#2003) to collect de-identified data (with patient consent). The surgical procedures involved left frontal cavernoma, right sphenoid wing meningioma, left petroclival chondrosarcoma, brainstem cavernoma and right sphenoorbital meningioma. These procedures tend to use up to five instruments simultaneously in the surgical field.

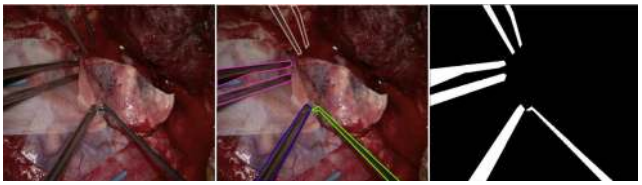

Figure 1. Manual annotation of a collected image frame to identify different Neurosurgical instruments (grasper, peach), (suction, blue), (curette, green) and (pickup, magenta).The binary segmentation ground truth from the annotation shows the tool pixels in white and background pixels in black.

Each video has a resolution of 720 x 480 px and runs at 29.97 frames per second. The images for the surgical tools dataset were collected every 14th frame to record information at an approximate rate of 2 frames per second. A total of 2400 images have been manually annotated with the help of neurosurgery fellows. Bounding polygons were used to generate segmentation ground truth for tool versus tissue and labels for tool-type. For the networks described in this paper, we had 850 annotated images (Fig. 1) available at the time of training, split into 700 training images and 150 test images with similar distributions. The dataset contains 8 different classes of instruments labeled for purposes of instance segmentation. We have additionally collected 6 more videos and will increase the size of our dataset and account for additional variability in surgical conditions.

Robotic Instrument Dataset: For robotic instruments, the MICCAI 2017 Endoscopic Vision Challenge Instrument Segmentation [1] dataset was used. Robotic instruments possess distinct articulated parts unlike neurosurgical instruments, namely rigid shaft, articulated wrist and grasper parts. The MICCAI Challenge supplied a dataset of stereo images taken from videos recorded *in vivo* using the da-Vinci surgical robot system.

3. Methods

In this work, three different deep architectures for binary segmentation were evaluated. The architectures incorporated an encoder-decoder structure, where the encoder network consists of successive convolutional layers, pooling layers and Rectified Linear Unit (ReLU) activations, capturing a compact feature map in an encoded latent representation. The pooling layers from the encoder are replaced with upsampling layers in the symmetric decoder network for recovering spatial information. The concatenation of higher resolution features from the downsampled path with the features in the upsampled section (Fig. 2) provides precise localization. The first architecture for the tool vs non-tool identification task is a Vanilla U-net [8]. This showed a vast improvement in performance compared to a heuristic-based baseline on both the datasets. The two other network architectures leverage transfer learning [12]. By modifying the encoder structure, different latent representations were obtained - one with a VGG16 encoder network[10] and another using a lightweight MobileNetV2[9] network, to improve runtime while not losing accuracy.

Training: For upsampling in the decoder network, we used bilinear interpolation for the Vanilla U-net and fractionally strided convolutions/transposed convolution with the for VGG16-UNet network. We updated our decoder for MobileUNet based on [11] to create a computationally efficient solution. The decoder uses a data-dependent upsampling technique called DUpsampling that incorporates better feature aggregation and downsamples the fused features to the lowest resolution before merging them.

The energy function used for training was:

$$L = H \log(J) \tag{1}$$

where, H is binary cross entropy loss function

$$H = -(ylog(p) + (1-y)log(1-p)) \tag{2}$$

and J is the Jaccard Index.

The VGG16-UNet and MobileUNet networks were pre-trained on the ImageNet dataset[5]. The dataset was augmented using translation, random horizontal and vertical flip, normalization, padding and random crop to increase its size and learn invariance properties. Each model was trained for 20 epochs, with a batch size of 4. The networks used Adam optimizer with an *alpha* of 0.0001. We used K-fold cross-validation during training. A threshold was set to binarize pixel probabilities following validation in each case. The networks were implemented using PyTorch, with evaluations on a machine with NVIDIA GTX 1080Ti.

4. Results

To show the relative performance of the neural networks with respect to the hand-crafted heuristic baseline

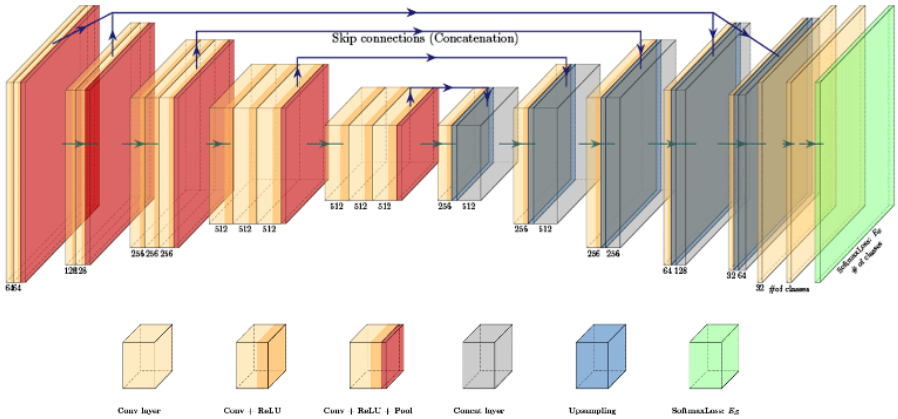

Figure 2. Encoder-Decoder architecture used for instrument segmentation. The number of channels is indicated below the box for a VGG encoder. The height of the box represents the feature map resolution.The yellow boxes represent convolution blocks, red represents pooling, blue represents upsampling and gray represents concatenation.

Table 1. Evaluation of performance - Dice Coefficient

Analysis	Neurosurgical Instruments		Robotic Instruments	
	Dice	IoU	Dice	IoU
Baseline Method	0.339	0.312	0.516	0.461
Vanilla U-net	0.6740	0.653	0.813	0.724
VGG16-UNet	0.736	0.7102	0.887	0.80
MobileUNet	0.769	0.748	-	-

(improved from method in [1]), the Dice coefficient and Intersection over Union (IoU) metrics were used to calculate quality of binary segmentation (Table 1). By incorporating a lighter MobileNetV2 network and modifying the downsampling technique, we were able to improve performance while generating a faster and lighter network to get best performance. The performance difference between the two datasets stem from the range of conditions in the datasets (more in NeuroID). We will incorporate a larger dataset for NeuroID in the next training iteration.

References

[1] Robotic instrument segmentation sub-challenge. https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/. Accessed: 2017-07-02.

[2] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis Herrera, Wenqi Li, Vladimir I. Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2017 robotic instrument segmentation challenge. *CoRR*, abs/1902.06426, 2019.

[3] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical Image Analysis*, 35:633 – 654, 2017.

[4] David Bouget, Rodrigo Benenson, Mohamed Omran, Laurent Riffaud, Bernt Schiele, and Pierre Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE Transactions on Medical Imaging*, 34(12):2603–2617, dec 2015.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[6] K. Jo, B. Choi, S. Choi, Y. Moon, and J. Choi. Automatic detection of hemorrhage and surgical instrument in laparoscopic surgery image. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1260–1263, Aug 2016.

[7] Austin Reiter, Peter K Allen, and Tao Zhao. Appearance learning for 3d tracking of robotic surgical tools. *The International Journal of Robotics Research*, 33(2):342–356, 2014.

[8] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[9] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[11] Zhi Tian, Tong He, Chunhua Shen, and Youliang Yan. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation. 2019.

[12] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *CoRR*, abs/1411.1792, 2014.