ONLINE SEGMENTATION AND TRACKING OF SURGICAL INSTRUMENTS FOR COMPUTER-AIDED SURGERY

NIVEDITHA KALAVAKONDA[1], ZEESHAN QAZI[2], BLAKE HANNAFORD[1], LALIGAM SEKHAR[2]
[1]UNIVERSITY OF WASHINGTON - SEATTLE, [2]HARBORVIEW MEDICAL CENTER

Overview

Computer-assisted surgery (CAS) and robotics reduce complications through the use of advanced instruments, control and visualization. Real-time tracking of surgical tools enable application of various computer-assisted techniques., such as augmented reality, to improve clinical outcomes.

We present our convolutional neural network designed for multi-task learning, to both track and perform semi-supervised instrument segmentation for real-time use during surgery.

Motivation

To perform data-driven surgical procedures, including robotic surgery, it is important to analyze instrument trajectories. Benefits include:

- Surgical workflow optimization
- Localization and relative pose estimation
- Instrument collision prediction/analysis
- Visual Servoing

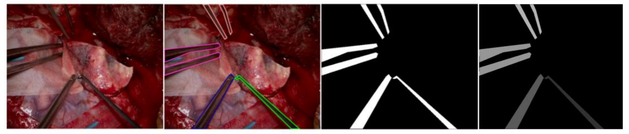

Fig1: Ground Truth Annotations from the NeuroID Dataset

Methods

Our method builds on top of Fully Convolutional Siamese Networks, trained on pairs of input video frames[4].
The Siamese network is trained on three tasks to establish correspondences between target object and candidate regions in new frames:

1. Learn measure of similarity between target object and multiple candidates in a sliding window approach [1]
2. Bounding box regression using Region Proposal Network [2]
3. Class-agnostic Binary Segmentation [3]

We obtain a multi-channel response map by processing an exemplar image(z) and a search image(x) through a Siamese network, yielding two depth-wise cross-correlated feature maps:

$$g_{th}(z, x) = f_{th}(z) * f_{th}(x)$$

For the box predictions, we use the L_1 loss for the box and cross-entropy losses for obtaining the score in the region proposal network.

To obtain the pixel-wise binary mask, we introduce an additional head, where mask prediction is a function of image to segment, x and target object, z

$$m_n = h_{phi}(g_{th}^n(z, x))$$

Schematic Illustration of Network

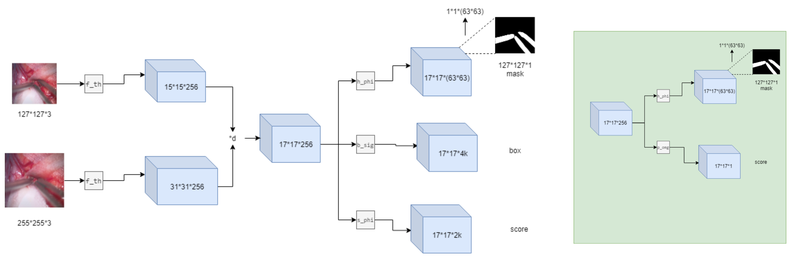

Fig2: Architecture showing the two- (L) and three-branch (R) variations we used for tracking and segmentation.

Mask prediction:

$$L_{mask}(th, phi) = \sum_n \left(\frac{1+y_n}{2wh} \sum_{ij} \log\left(1 + e^{-c_n^{ij} m_n^{ij}}\right)\right)$$

where y_n is the ground-truth label, c_n is the ground-truth mask.

The total loss is:

$$L_{total} = \lambda_1 L_{mask} + \lambda_2 L_{score} + \lambda_3 L_{box}$$

In the two-variant head, we do not include the loss function for the box.

Backbone Architecture (f_th)

block	***examplar* output size**	***search* output size**	***backbone***
Conv1	61×61	125×125	7×7, 64, stride 2
Conv2_x	31×31	63×63	3×3 max pool, stride 2 $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3_x	15×15	31×31	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 64 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Conv4_x	15×15	31×31	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
adjust	15×15	31×31	1×1, 256
xcorr	17×17		Depth-wise

Results

For the backbone, we make use of the ResNet-ro architecture (until the final convolution layer of the 4th stage). All models are trained using COCO, ImageNet-VID and YouTube-VOS, and fine-tuned on the surgical instrument dataset we have generated, called NeuroID.

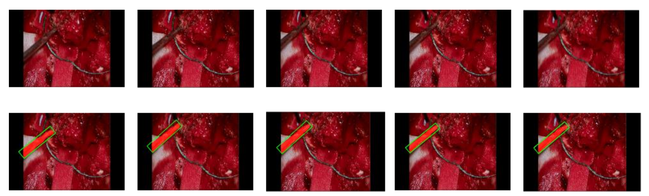

Fig3: Results on suction instrument across frames. The model refines its segmentation between frames and improves on its prediction.

For generating a bounding box from the segmentation map, we used a rotated mimum bounding rectangle. We also tested with the optimization strategy for automatic bounding box in [5].

The network performed at an average of 57 frames per second (fps) on an NVIDIA TX 2060 GPU. The feature extractor used a large amount of this processing time.

Evaluation on VOT-2016: mIOU for the network was 67.15% using the Minimum bounding rectangle evaluation and 71.68% while using the optimization strategy from [5] for the BBox.

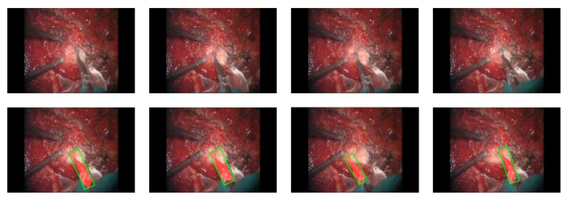

Fig4: The network has some limitations wrt segmenting instruments which are occluded by the surgeon's hand. They, however, can perform well under occlusion from another instrument.

Proposed improvements:

1. We are currently working on improving the network performance under heavy occlusion conditions
2. Another challenge is maintaining the frame rate when there are multiple instruments in the frame, which we hope to alleviate by using a lighter backbone network

References

[1] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In European Conference on Computer Vision workshops, 2016.
[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, 2015.
[3] P. O. Pinheiro, R. Collobert, and P. Dollar. Learning to seg- ´ ment object candidates. In Advances in Neural Information Processing Systems, 2015.
[4] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In CVPR. IEEE, 2019.
[5] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, L. Cehovin, T. Vojir, G. Hager, A. Lukezic, G. Fernandez, et al. The visual object tracking vot2016 challenge results. In European Conference on Computer Vision, 2016.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON