Endourology Society Summer Scholarship 2024

Student: Nathan Schuler, University of Illinois College of Medicine, Chicago; Class of 2027

Mentor: Ahmed Ghazi, MD, Associate Professor of Urology, Johns Hopkins Brady Urological Institute, Director of Robotic Surgery, Johns Hopkins Hospital

Development of a machine learning-driven, automatic video gesture labeling system for use in Nerve sparing Robot-assisted Radical Prostatectomy

Introduction

Current methods of surgical evaluation are centered around global metrics, assessed through video review of surgical procedures. Of these, the current standard is GEARS which can be applied to general robotic aptitude[1], and DART, which was recently developed specifically for soft tissue dissection, and then validation with respect to clinical outcomes[2]. Recent work has identified Gestures – granular individual tasks surgeons complete throughout the procedure like cuts, dissection, and retraction – as a possible source of metrics that can be applied in an objective manner to evaluate surgery. Previous work by Ma et al has shown that surgical gesture analysis can be used to predict clinical outcomes with an ROC AUC of 0.83[3]. Similarly, analysis of kinematics data from the DaVinci console has similarly shown a link between surgeon movements and clinical outcomes in radical prostatectomy procedures[4]. Reversing the question from looking at clinical outcomes to the surgeons themselves, we have previously used surgical gestures and low-level machine learning to assess the expertise of surgeons themselves using high or low procedure specific caseload as the metric for expertise, achieving a ROC AUC of 0.96[5].

Transformer-based neural network models have formed the basis for natural language processing in machine learning, but also have shown promise when utilized for computer vision tasks[6]. This is due to their ability to utilize positional encoding to take into account context beyond the single frame of observation[7]. With this in mind we hypothesized that a transformer-based architecture might be an ideal use case for event recognition in surgery. Given objectivity and effectiveness of gesture analysis within its limited body of work, we believe that if we are able to create a system that is automatically able to identify these movements, there is potential for a near-immediate postoperative evaluation of the technical skill utilized by the surgeon throughout the procedure, for supplementation with other clinical factors as well as a comparison to the individual surgeon baseline.

Methods

Our previous work generated a total of 50 surgical videos within a realistic nerve sparing robot assisted radical prostatectomy simulation. These videos were captured in 30 frames per second, and annotated using a hierarchical task analysis-derived standardized methodology for surgical gestures as previously described by Vedula et al[8]. From this we established a database of over 35,000 surgical gestures matched to timestamps within recorded video. Using python we developed a customizable script capable of separating video into discrete, user-defined durations, matched with an individual gesture associated with each video. Using the keras API with the TensorFlow module in python, we constructed a transformer-based computer vision framework to accomplish two functions. First, we used the ViVIT pre-trained image feature extraction transformers to extract features from every video in the database, for the purpose of training our own model. Concurrently, gesture vectors were constructed from the timestamped video labels. Subsequently, we trained our naïve transformer model on the extracted features and then tested the model with a 90:10 train:test split. We assessed the model across various video resolutions and

numbers of heads within the transformer model (number of consecutive video frames utilized for training and prediction). After predicting gestures in video, confusion matrices were generated encompassing all ten gestures, along with an assessment of model accuracy in the form of their ability to correctly predict the gesture within the corresponding video segment. Also included was a distribution of the gestures that were predicted by the transformer model in cases where the prediction was incorrect.

Preliminary Results:

A total of 38,416 Surgical Gestures were included in training the transformer model, encompassing ten distinct actions (Table 1)

Table 1: Total Numbers of Surgical Gesture across all Participating surgeons

Blunt Dissection	1,889
Clip	71
Cut	9,948
Sharp Dissection	9,919
Left Dynamic Tension Apply	5,490
Left Dynamic Tension Release	5,011
Right Dynamic Tension Apply	2,009
Right Dynamic Tension Release	1,770
Static Tension Apply	1,302
Static Tension Release	1,007
Total:	38,416

Table 1: Summary table of surgical gestures included in the video segmentation, training, and testing cycles.

The prediction model overall achieved low accuracy, with a maximum whole-model accuracy of 36% with a ROC AUC of 0.52, achieved when utilizing 16 heads – analogous to a 0.53 second window for video classification. (Figure 1).

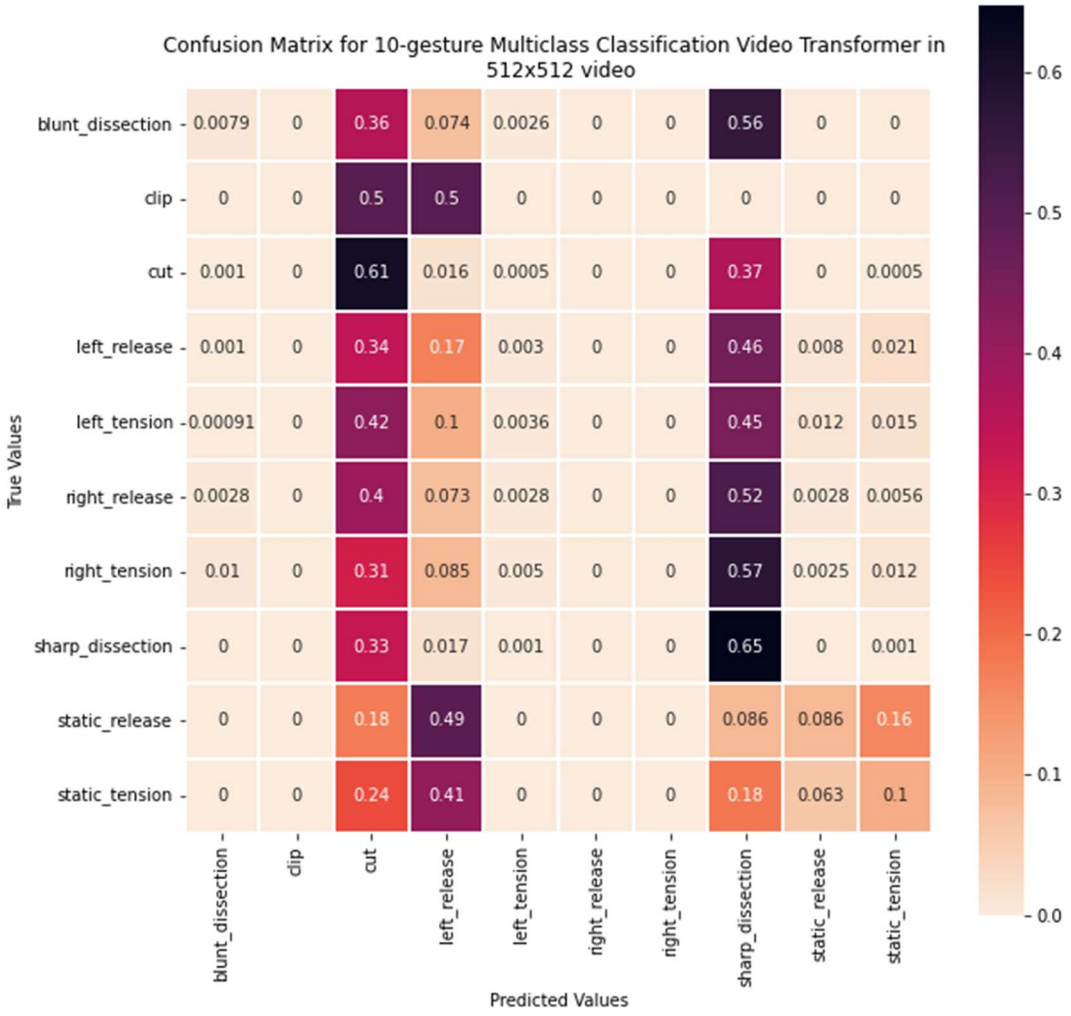

Figure 1: Confusion matrix representation of model predictions versus actual labels, utilizing 1-second duration videos downscaled to 512x512 resolution with 512 features extracted per frame, and with 16 headed transformer network.

The model showed positive predictive capabilities greater than 50% for two gestures – Cut (61%) and Sharp Dissection (65%) (Figure 1). Blunt dissection (0.7%), clip (0%), left dynamic tension release (17%), left dynamic tension application (0.4%) right dynamic tension application (0%) right dynamic tension release (0%) static tension application(10%), and static tension release (8.6%) were highly likely to be mislabeled as either cuts or sharp dissections (Figure 1). Less frequently, predictions were incorrectly assigned to left dynamic tension release (Figure 1).

Preliminary Conclusions

We have achieved our first aim of developing a pipeline for generating and manipulating a database of basic gestures matched to surgical video. We have subsequently achieved our second aim of creating a transformer-based computer vision model that is able to take inputs extracted from our generated database, train, and predict the movements that are occurring. The model we produced achieved low accuracy in selecting the primary gesture occurring within a one-second duration video, achieving a maximum of 36% accuracy with 0.52 ROC AUC. These results are weak, achieving predictive value only marginally better than chance alone as a whole. One notable result was that the model did show an ability to predict true cuts and sharp dissections at higher rates than alternatives, however the inability to predict

movements outside of these two with any real significance outweighs this small positive. The work we have completed to this point has several weaknesses that became apparent throughout the analysis. First, an inherent weakness of training on video segments exists wherein it is possible that a discrete video segment may include multiple gestures depending on the speed at which the surgeon is working, and the subsection of the dissection task they are performing. The presence of complete, or incomplete secondary gestures may serve as a substantial confounding factor that detracts from the ability of the model to assess the single gesture that is the focus of the video. Furthermore, the methodology from Ma et al[3] using gestures and demographic data to predict clinical outcome data included higher complexity data structures, where the extracted features were then appended to the raw video, and subsequently used for training their model. This would be possible within our current experimental structure; however, it would not address the need to differentially represent potential confounding gestures.

Future Direction

Given the low accuracy of the model, with minimal improvement upon modification of several hyperparameters, we realize a need to alter our methodology. We have begun feature extraction on the full-length surgical videos on a frame-by-frame basis, and additionally reconstructed the gesture data to be continuous throughout the duration of the video as well and matched to the corresponding time points within the full video. While this may increase the computational resources required, we believe it has potential to produce a more capable and robust model. We plan to continue working toward implementing these changes and reassessing model predictive capabilities accordingly.

References

1. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *J Urol*. 2012;187(1):247-252. doi:10.1016/j.juro.2011.09.032

2. Vanstrum EB, Ma R, Maya-Silva J, Sanford D, Nguyen JH, Lei X, Chevinksy M, Ghoreifi A, Han J, Polotti CF, Powers R, Yip W, Zhang M, Aron M, Collins J, Daneshmand S, Davis JW, Desai MM, Gerjy R, Goh AC, Hu JC, Kimmig R, Lendvay TS, Porter J, Sotelo R, Sundaram CP, Cen S, Gill IS, Hung AJ. Development and Validation of an Objective Scoring Tool to Evaluate Surgical Dissection: Dissection Assessment for Robotic Technique (DART). *Urol Pract*. 2021;8(5):596-604. doi:10.1097/UPJ.0000000000000246

3. Ma R, Ramaswamy A, Xu J, Trinh L, Kiyasseh D, Chu TN, Wong EY, Lee RS, Rodriguez I, DeMeo G, Desai A, Otiato MX, Roberts SI, Nguyen JH, Laca J, Liu Y, Urbanova K, Wagner C, Anandkumar A, Hu JC, Hung AJ. Surgical gestures as a method to quantify surgical performance and predict patient outcomes. *Npj Digit Med*. 2022;5(1):187. doi:10.1038/s41746-022-00738-y

4. Hung AJ, Chen J, Che Z, Nilanon T, Jarc A, Titus M, Oh PJ, Gill IS, Liu Y. Utilizing Machine Learning and Automated Performance Metrics to Evaluate Robot-Assisted Radical Prostatectomy Performance and Predict Outcomes. *J Endourol*. 2018;32(5):438-444. doi:10.1089/end.2018.0035

5. Schuler N, Shepard L, Saxton A, Russo J, Johnston D, Saba P, Holler T, Smith A, Kulason S, Yee A, Ghazi A. Predicting Surgical Experience After Robotic Nerve-sparing Radical Prostatectomy Simulation Using a Machine Learning–based Multimodal Analysis of Objective Performance Metrics. *Urol Pract*. 2023;10(5):447-455. doi:10.1097/UPJ.0000000000000426

6. Parvaiz A, Khalid MA, Zafar R, Ameer H, Ali M, Fraz MM. Vision Transformers in medical computer vision—A contemplative retrospection. *Eng Appl Artif Intell*. 2023;122:106126. doi:10.1016/j.engappai.2023.106126

7. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. Published online 2017. doi:10.48550/ARXIV.1706.03762

8. Vedula SS, Malpani AO, Tao L, Chen G, Gao Y, Poddar P, Ahmidi N, Paxton C, Vidal R, Khudanpur S, Hager GD, Chen CCG. Analysis of the Structure of Surgical Activity for a Suturing and Knot-Tying Task. *PloS One*. 2016;11(3):e0149174. doi:10.1371/journal.pone.0149174