EN 601.496 Computer Integrated Surgery II

Surgical Skill Analysis with Virtual Reality

Final Report

05.01.2022

Aditya Khandeshi, Liza Naydanova, Alexandra Szewc

Mentors: Max Li, Adnan Munawar, Dr. Danielle Trakimas, Nimesh Nagururu, Dr. Creighton, Dr. Unberath, Dr. Taylor

I. Background

Evidence has shown that higher-volume surgeons with superior technical skills yield better patient outcomes.[1-3] However, many factors such as restrictions of duty hours and an increase in the overall complexity of the procedures that trainees need to achieve competence in have reduced the number of autonomous training experiences available to surgical trainees. As a result, it is necessary to implement accurate and effective methods by which trainees can objectively be evaluated on their competencies. Currently, the widely-used methods for evaluating surgical skill include the Objective Structured Assessments of Technical Skills (OSATS) and the Society for Improving Medical Professional Learning platform. However, some limitations of these evaluations include bias from the evaluator, as well as poor intra-evaluator reliability.[4-6] Additionally, correctly implementing OSATS is time consuming and resource intensive, incentivizing the discover of more efficient assessment techniques.[7] Whereas other fields of surgery have made progress in testing technical skill in some sort of simulated environment, the field of Otolaryngology Head and Neck Surgery (OHNS) has not yet made any significant leaps in objectively assessing technical skills in simulated environments.

II. Problem

The problem addressed by our project is the development of a method for objective, technical assessment of the skill of surgeons performing mastoidectomies. Mastoidectomies are procedures that aim to remove cells either from unwanted growth or infection in the mastoid bone behind the ear. This procedure requires drilling of the temporal bone, which must be done carefully in order to avoid damaging any of the numerous vital structures nearby, and so ensuring that trainees are sufficiently competent is paramount for such a procedure. Advances in virtual reality have allowed for the creation of temporal bone simulators, many of which provide haptic feedback and stereoscopic vision to create an immersive and realistic environment for trainees. However, many automated metrics that have been generated have many limitations, including a lack of information on user specific errors and a lack of validation across multiple temporal bone specimens.[8] Since these such simulators show great promise for use in OHNS technical assessment, the lack of any applicable automated metrics show that there is still a need to both define and measure objective metrics of technical skill in this field in order to analyze surgical competency, which is what this project aims to address.

III. Approach

Our project mentors provided a software called Asynchronous Multibody Framework (AMBF), which was used along with a PHANToM Omni Haptic Device to simulate and then record mastoidectomy procedures performed by different clinicians and develop an algorithm for evaluating skill level. In order to develop an algorithm for evaluating skill level, features in the data connected to surgical skill level were identified via a literature review of technical papers Azari et al. (2019) and Wijewickrema et al. (2015), among other resources. We then created scripts to extract the pertinent features from the data collection pipeline reported by the AMBF temporal drilling simulator. This pipeline stores its reported data in the form of an hdf5 file, which were then parsed by our scripts in Python. In order to ensure that the various metrics selected could accurately be calculated, the simulation data stream needed to be modified to report all necessary measurements over the course of a simulation run. Prior to the literature review, the data stream reported the position of the drill with both the location of the tip and its

orientation. We had timestamps indicating the time at which measurements were taken and stereoscopic RGB images to reconstruct a video of the procedure. At each timestamp, if a voxel in the temporal bone volume was removed, that specific voxel's location was reported. Depth and segmentation maps were also provided, along with the pose of the temporal bone volume as well as instances where the size of the drill burr was changed. Based on the selected features, force data from the PHANToM Omni device needed to be added to the data collection stream. Finally, the feature extraction procedure had to be validated with the data collected by the modified pipeline to ensure that future skill assessment would be interpretable for the following extracted data motivated by our critical review:

Extraction of Force From Data Collection Stream

The volumetric drilling simulator data collection stream relies on a set of ROS messages that store relevant data at each timestamp of the collection, and are then subscribed to. In order to collect the force data from the PHANToM Omni, a new ROS message was created that captured the wrench, a ROS geometric message storing both the linear and angular components of force at a timestamp, setting a publisher for the message, calling the message in the simulator to store measurements at necessary timestamps, and creating a subscriber for this publisher in the final data recording file.

Surgical Strokes Segmentation

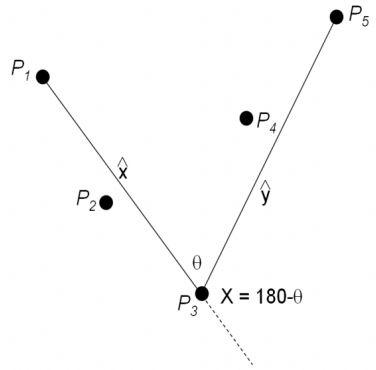
The analysis of distinct surgical motions is integral to the extraction of features for skill analysis.[9] Since the recorded data appears in the format of a stream of drill positions over time, strokes could be identified as a set of points representing a continuous drilling motion by defining the end of a stroke as point reached when the direction of the drill trajectory changes abruptly. The stream of positional data was represented by the time series T in (1). Then, the stream was separated into micro-stroke subsets as described by (2). This could be taken a step further through the use of some classification function, F_c, which separates strokes by designating some position in the data stream as the last in a given stroke, as shown in (3). F_c then has to be defined using some information-rich attribute of each point such that deviation of a point's value by a standard deviation from the mean constitutes a positive classification. Such an information-rich attribute, X_P in (4), is defined over an interval of k=6 points, as suggested by the literature. Finally, the k-cosine metric, as demonstrated in (5), was used to define X_P for each point in order to classify stroke boundaries.[10]

$$T = \{P_j = (x_j, y_j, z_j) | j = 1, 2, 3, \ldots, m\}; \quad (1)$$

$$T = \{S_i = \{(x_{ij}, y_{ij}, z_{ij}) | j = 1, 2, 3, \ldots, m\} | i = 1, 2, 3, \ldots, n\} \quad (2)$$

$$F_c(X_{P_b}) = \begin{Bmatrix} 0: & P_j \in \mathrm{S}_i, & P_{j+1} \in \mathrm{S}_i \\ 1: & P_j \in \mathrm{S}_i, & P_{j+1} \in \mathrm{S}_{i+1} \end{Bmatrix} \quad (3)$$

$$X_{P_j} = F_b(P_{j-k}, P_j, P_{j+k}) \quad (4)$$



k-cos (5)

Surgical Phase Segmentation

Discussion with our mentor Dr. Danielle Trakimas revealed that it would be important to distinguish between the phases of the surgical procedure. To do this, it would be necessary to manually split the timestamps of each simulator recording into the different phases of the temporal drilling procedure and for a set of metrics to be extracted for each phase for a deeper comparison between expert and trainee competence. This means that a trained expert—either Dr. Trakimas or Dr. Creighton—must label specific timestamps that indicate the start of a new phase of the drilling procedure.

Using these data augmentation methods, the following features were then extracted and validated:

Drill Velocity, Acceleration and Jerk

Drill kinematics, are extremely useful in determining drill technique and can vary between experts and trainees.[7] Thus for each segmented stroke, the average velocity, acceleration and jerk were calculated using the positional data of the drill tip. The numerical derivatives of the position data were calculated in order to obtain velocity information for the data, which was then used to calculate the acceleration information as well as the magnitude of the average velocity of the stroke. The numerical derivatives of the acceleration data were then computed in order to calculate the average jerk. The mean, median, and maximum values for each kinematic metric were then reported across all strokes for each phase of the procedure.

Spatiotemporal Curvature

Similar to the basic drill kinematics, spatiotemporal curvature is also another kinematic metric that is useful for assessing drill technique.[7] The numerical first and second derivatives of the position data were calculated and were then used in order to calculate the curvature using the formula to the right.[11] The mean, median, and maximum values were then reported across all strokes for each phase of the procedure.

$$\kappa(t) = \frac{\|\mathbf{r}'(t) \times \mathbf{r}''(t)\|}{\|\mathbf{r}'(t)\|^3}$$

Force

The force vector data reported by the PHANToM Omni Haptic Device was used to calculate the average magnitude of the force exerted by the drill tip over the interval of each surgical stroke. The mean, median, and maximum stroke force values were then reported across all strokes for each phase of the procedure.

Stroke Length

Using the segmented stroke boundaries and the positional data of the drill tip, the total distance traversed by the drill across each stroke was estimated by summing the distance between each sequential position recording. The mean, median, and maximum stroke force length were then reported across strokes for each phase of the procedure.

Bone Removal Rate

Using the segmented stroke boundaries and the positional data of the drill tip, the total number of voxels removed during the time interval of each stroke was extracted and divided by the time interval. The mean, median, and maximum stroke force length were then reported across strokes and for each phase of the procedure.

Procedure Duration

The duration of the procedure was defined as the difference between the time the first and final voxels of the model were removed.

Drill Orientation

Drill orientation with respect to the plane of the bone at any given time was computed using force data across time points of the procedure. For all time points with nonzero force recordings, the force vector, which is normal to the plane of the bone, was used to compute an angle between the drill tool and the bone plane. The mean, median, and maximum stroke force length were then reported across strokes and for each phase of the procedure.

Appropriate/Inappropriate Removal of Bone

Appropriate and inappropriate removal of bone was characterized by voxels of the temporal bone volume where at least 5 of 6 experts drilled bone and did not drill bone respectively. In order to calculate this, 6 expert drilling simulations would need to be conducted, and a label map of the appropriately removed voxels of the bone volume created. Raw counts as well as percentages of appropriate voxels removed are then reported. Similarly, inappropriate removal of bone was characterized by voxels of the temporal bone volume that had at most a 17% chance of removal by experts. In order to calculate this, the same 6 expert drilling simulations would need to be collected, but this time specified voxels would be those that were removed in at most 1 of the 6 runs. As with appropriate removal, raw counts as well as percentages are reported.

Finally, a baseline random forest machine learning model was defined to be used with acquired training data on which the aforementioned features were extracted. The implementation of this baseline will guide the further development of skills analysis models, such as those based on convolutional neural network algorithms.

IV. Results

The features extracted in the aforementioned sections were validated on individual data collection trials to ensure that appropriate and accurate values were returned. This was accomplished through the collection of force, positional, and voxel removal data over a series of specialized trials to target edge cases for each feature. Specialized trials included those which varied:

Surgical Stroke Count

Data collection trials with 0, 3, and 9 strokes were each recorded to validate the ability of our feature extraction code to adequately segment strokes.

Drill Kinematics

Data collection trials with slow, constant movement and rapid, variable-speed movements were recorded and quantitatively assessed relative to one another to ensure that velocity, acceleration, and jerk are being recorded appropriately.

Stroke Force and Bone Removal Rate

In order to validate both of these two features, trials for data collection were conducted with no bone removed, some bone removed slowly, and with bone rapidly removed at great force. The extracted features were then quantitatively assessed relative to one another to validate whether force and bone removal reate appeared correlated.

Stroke Length and Spatiotemporal Curvature

For the validation stroke length and spatiotemoral curvature, trials with small straight strokes and long straight strokes were recorded and quantitatively assessed relative to long, curved-stroke trials. Validation was passed if curvature was greater for the latter trials than for the first and if stroke length was reported as increasing in the order that these trials are listed.

Procedure Duration

Reported procedure duration was validated against a manual recording of the procedure duration.

Varied Drill Angle

Data collection trials with 45 degree, 90 degree, and randomly-angled strokes were recorded to validate that the drill orientation feature extraction procedure outputs reasonable angles between the drill and the plane of the bone. These angles were quantitatively assessed relative to one another and the expected values to ensure that velocity, acceleration, and jerk are being recorded appropriately.

In addition to the above mentioned validation trials for feature extraction, all features were additionally validated on all other validation trials to ensure there were no unexpected deviations from expected behavior.

V. Significance

The results described above are significant towards the solution of the problem we identified above: the development of a method for objective, technical assessment of the skill of surgeons performing mastoidectomies. By researching relevant literature to identify pertinent features, extracting the identified features, and validating the accuracy of the extracted features in addition to the baseline implementation of a machine learning model to analyze skill, we have laid the foundation for the development of an accurate and interpretable model for skill assessment. From the results reported and the documentation provided, the collection of additional data from surgical experts and trainees will generate a preliminary skill assessment model that will be scalable with continued data collection.

VI. Management Summary

Group members worked together evenly on the project plan, critical review, and project checkpoint presentations. Much of the simulator setup on the group's work laptop was performed by Liza. Aditya worked on integrating the PHANToM Omni and HTC Vive VR Headset external hardware into the simulator. Alexandra was responsible for documenting progress and managing and updating the group's project Wiki and Github to maintain scripts and collected simulator data. Aditya added force measurement and eye-tracking to the data collection pipeline, whereas Alexandra and Liza were responsible for research and analysis for the critical review and feature extraction. All group members coordinated the selection and computation of features from recorded data. Aditya and Alexandra were then responsible for implementing feature computation procedures and generating unit tests in order to validate their accuracy. Documentation of code, simulator setup, feature extraction, and feature extraction validation was maintained by Aditya and Alexandra. Finally, Alexandra was responsible for model implementation, while Aditya oversaw data collection for validation and model training.

The primary aims of the project for the semester were the selection and documentation of skill-assessment metrics, the creation of scripts to extract these metrics from the modified AMBF simulator data collection stream, and feature validation alongside both skilled and unskilled data recording for script validation as well as guiding the creation of a machine learning model to predict surgical skill. These aims deviated slightly from our initial aims, which focused more heavily on data collection in lieu of feature extraction, which we found to be much more important for the future success in solving our identified problem. Overall, the group was able to ensure the completion of all primary goals as well as document them in addition to ongoing progress. However, due to unforeseen health complications for a group member, we were presented with significant delays in progress, which ultimately led to our goals requiring more time than originally planned. While ultimately making up for some of the lost time, our group was unable to make as significant headway into the final skill assessment as was planned, with the majority of the focus being placed into the validation of feature extraction and collection of as much data as was possible despite the setbacks in the timeline. However, the extra time that was taken for feature validation will ensure that the metrics extracted from the data collection pipeline will be accurate and reliable when the time comes to train our network's architecture.

The group also has future plans to continue work on the project. Alexandra and Aditya will be continuing with the project over the summer, where the next steps would be to finalize the design of our metric extraction to a neural network workflow and train the network with collected data from trainee and expert OHNS surgeons. The end goal following training and validation of the network's predictions is to publish our results in a conference paper.

Overall, the project has given an insightful experience into the various different moving parts that go into virtual-reality based software. The majority of the project's software was performed on a Linux system, so learning how to handle the installation and management of both the VR headset and eye trackers was an important skill to learn. Even more important, understanding the AMBF and volumetric drilling code base, which seamlessly combined ROS, C++, and Python to create a functioning simulator and data collection pipeline was extremely enlightening. Understanding how to modify or add to the data collection pipeline was another skill learned during the project, which was a valuable experience given that nobody in the group had any prior experience working with ROS. Ultimately, these newly gained skills will continue to aid in the group's progress as the semester closes and as we continue our progress throughout the summer and upcoming academic year.

VII. Technical Appendices

A. Documentation Sources, Access, and Lifetime

Our group used several platforms to create and modify our project materials. We have consolidated everything to be accessed in both modifiable and persistent formats using GitHub and our assigned CIIS Project Wiki page. This was done to ensure constant availability of the content used, produced, and generated by our project for the foreseeable future, and beyond our time at JHU. These management platforms can be found at:

- Project GitHub:

  https://github.com/aszewc1/vr-skill-analysis

- Project Wiki Page:

  https://ciis.lcsr.jhu.edu/doku.php?id=courses%3A456%3A2022%3Aprojects%3A456-2022-08%3Aproject-08

The link to this group's repository as well as all references used throughout the duration of the project can also be found on the Project Wiki. The Project Wiki also contains PDFs of all course reports and presentation slides throughout the semester, typically stored as both a modifiable, collaborative source link as well as a persisting, export format such as pdf.

B. Code Access and Documentation

All code, documentation of code including proper usage and purpose, and collected data used for validation is managed in a GitHub repository at the Project GitHub listed above.

C. Simulator Setup Documentation

A documented procedure for setting up and running the simulator is included in our Project GitHub listed above under the title "AMBF Simulator and Phantom Omni Setup."

Additionally, the modified data collection pipeline code with includes force calculation based on the initial volumetric drilling code provided by our mentors can also be found in our Project GitHub listed above.

VIII. References

[1] Al-Qurayshi Z, Robins R, Hauch A, Randolph GW, Kandil E. Association of Surgeon Volume With Outcomes and Cost Savings Following Thyroidectomy: A National Forecast. JAMA Otolaryngol Head Neck Surg 2016; 142:32-39.

[2] Kim HI, Kim TH, Choe JHet al. Surgeon volume and prognosis of patients with advanced papillary thyroid cancer and lateral nodal metastasis. Br J Surg 2018; 105:270-278.

[3] Birkmeyer JD, Finlayson EV, Birkmeyer CM. Volume standards for high-risk surgical procedures: potential benefits of the Leapfrog initiative. Surgery 2001; 130:415-422.

[4] Meyerson SL, Sternbach JM, Zwischenberger JB, Bender EM. The Effect of Gender on Resident Autonomy in the Operating room. J Surg Educ 2017; 74:e111-e118.

[5] Gerull KM, Loe M, Seiler K, McAllister J, Salles A. Assessing gender bias in qualitative evaluations of surgical residents. Am J Surg 2019; 217:306-313.

[6] Sethia R, Kerwin TF, Wiet GJ. Performance Assessment for Mastoidectomy. Otolaryngol Head Neck Surg 2017; 156:61-69.

[7] Azari, D. P., Frasier, L. L., Quamme, S., Greenberg, C. C., Pugh, C. M., Greenberg, J. A., & Radwin, R. G. (2019). Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating. *Annals of surgery*, *269*(3), 574–581. https://doi.org/10.1097/SLA.0000000000002478

[8] Wijewickrema S, Talks BJ, Lamtara J, Gerard JM, O'Leary S. Automated assessment of cortical mastoidectomy performance in virtual reality. Clin Otolaryngol 2021; 46:961-968.

[9] Wijewickrema, S., Piromchai, P., Zhou, Y., Ioannou, I., Bailey, J., Kennedy, G., & O'Leary, S. (2015). Developing effective automated feedback in temporal bone surgery simulation. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*, *152*(6), 1082–1088. https://doi.org/10.1177/0194599815570880

[10] Hall, R, Rathod, H, Maiorca, M. Towards haptic performance analysis using K-metrics. Haptic and Audio Interaction Design. 2008;50-59.

[11] Rao, C., Yilmaz, A. & Shah, M. View-Invariant Representation and Recognition of Actions. International Journal of Computer Vision 50, 203–226 (2002). https://doi.org/10.1023/A:1020350100748