EN 601.656 Computer Integrated Surgery II

Surgical Skill Analysis with Virtual Reality

Background Literature Critical Review

03.10.2022

Aditya Khandeshi, Liza Naydanova, Alexandra Szewc

Mentors: Max Li, Adnan Munawar, Dr. Creighton, Dr. Unberath, Dr. Taylor

I. Introduction to Project

The goal of our project is to develop an objective, technical assessment of the skill of surgeons performing mastoidectomies. 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, these have limitations due to bias and intra-evaluator reliability. We plan to use a virtual reality framework to 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 must be identified. To explore possibilities for informative features, we examined Azari et al. (2019) and Wijewickrema et al. (2015).

II. Critical Review of Azari et al. (2019)

A. Overview

This paper from the University of Wisconsin-Madison used computer vision to predict expert performance ratings from surgeon hand motions as they performed tying and suturing tasks. Open surgeries were videoed and surgeon hands were tracked. A panel of expert attendings rated video clips, and empirical models were developed to predict the attendings' ratings based on recorded hand motion. It was found that the developed algorithm consistently predicted the expert consensus rating of individual tasks more objectively and reliably than individual assessment by surgical experts.

B. Relevant Background

Since correctly implementing OSATS is time consuming and resource intensive, there has been a push to discover more efficient assessment techniques. Motion capture and tracking of surgeon hand movements have the potential to achieve this goal. However, many methods depend on space-consuming systems and markers or sensors placed on the surgeon's hands. The authors focused on video motion capture of the surgeon's hands, as that method can be implemented to be a non-invasive and scalable means of observing surgical motion. This study builds on previous work in developing the system and kinematic models of hand motion to compare expert ratings of surgical skills to kinematic measures of surgeon hand motions.

C. Methods

Participants. The study involved 9 surgeons (6 attendings, 3 residents) doing 16 surgical cases.

Video Selection. Recorded videos were reviewed using Multimedia Video Task Analysis (MVTA) software, and a researcher familiar with operative technical tasks categorized the videos in the software for segments of tying/suturing tasks where hands were clearly visible for a minimum of 5 seconds.

Rating Scales. The authors used subjective visual-analog rating scales, adapted from the OSATS motion scales, for evaluating motion economy, fluidity of motion, and tissue handling. Fluidity of motion measured "hesitance, pauses or changes in direction and "resets"" (Azari et al.). Tissue handling refers to how appropriate the surgeon's force is when working with tissue, and varies with the type of tissue. Motion economy refers to how efficient the movement of the surgeon is.

After collecting and labeling data, a consensus panel of 3 expert surgeons viewed the clips in random order and independently rated hand motion based on the scales. Ratings were announced and discrepancies discussed until consensus was reached.

Motion Tracking. A custom video tracking software was used to trace a region of interest (ROI) across video frames.

Calibration. Visible measurements of the hands were used to calibrate each recorded clip from pixels into millimeters. Proximal interphalangeal joint breadth was scaled to the population of males or females depending on the sex of the surgeon.

Variable Selection. Instantaneous displacement, speed, acceleration, jerk, and spatiotemporal curvature were quantified based on the tracked record of the ROI. The data was smoothed via a second order Butterworth filter and an FFT was applied to remove cyclic and repeated motion patterns.

Modeling Process. A set of linear regression models was used to test whether kinematic features could predict expert ratings across the motion scales. Subsets of predictor variables were selected to run regression analysis. Authors predicted that motion economy would correspond to how fast surgeons moved in an area, fluidity would correspond to how many changes in speed there were, and tissue handling would be sensitive to any sudden changes in direction.

Validation. To assess internal validity, the sum of squared errors (SSE) was compared to the leave-one-out predicted residual sum of squares (PRESS).

D. Results

Video Data. There were a total of 103 video clips with a mean length of 11.72 seconds recorded. Hand tracking data provided over 1500 kinematic variables.

Task Expert Rating Scales. Ratings for suturing tasks had a mean of 5.91 and standard deviation of 1.62, and ratings for tying tasks had a mean of 7.12, standard deviation of 1.10, and were skewed towards the higher values.

Prediction Models of Expert Ratings. For suturing tasks, models for fluidity of motion (slope=0.86, R^2=0.86) and motion economy (slope=0.89, R^2=0.88) had the best predictions, performing moderately better than tissue handling (slope=0.76, R^2=0.69). For tying tasks, models of motion economy (slope=0.65, R^2=0.64) performed better than tissue handling (slope=0.53, R^2=0.52) and fluidity of motion (slope=0.54, R^2=0.54).

Prediction Model Validity. Linear prediction models were validated by comparing the error between PRESS and SSE. Motion economy models had the smallest average error for suturing ratings overall (0.27). Fluidity of motion models had overall error for suturing and tying (0.21, 0.11). Tissue handling had the highest error for both suturing and tying tasks (0.52, 0.26).

E. Evaluation

The importance of this paper is that it took a step towards creating more objective, reproducible and accessible assessments of surgical skill, obtained from video data. The prediction models developed can also be used to create an automatic feedback system for training settings. This paper is relevant to our project due to some of the kinematic metrics of

skill it defines. While the metrics evaluated skill during suturing and tying tasks, they were quite generally adapted from OSATS, and are likely to be applicable in other types of surgical procedures.

The authors discussed certain limitations to their work. First, the range of scores was limited by the videos that were available, which led to a small variance, making tasks more difficult to predict. Second, surgical context could not be addressed by the rating scales. Finally, the authors noted that patient and procedure outcomes were not taken into account, so an error and a successfully completed procedure look identical. We noted some possible issues with the paper as well. First, when consensus discussions happen between expert evaluators, it is possible that groupthink may contribute to outlier scores being dismissed without good cause. Second, more clarity about what features were included in each of the models, or if all features were used every time would have been helpful. Finally, all measurements were done assuming the surgeon's proximal interphalangeal joint breadth was the mean for the gender. It brings into question how much/if any of the results would be different if the actual joint breadth was used.

Next steps for this work would be to collect videos of specific clinically simulated scenarios, so that variety and experience could better be controlled. This would serve to make distributions less skewed. In addition, information about patient and procedure outcomes can be incorporated into these models to distinguish purposeful movements and errors.

III. Critical Review of Wijewickrema et al. (2015)

A. Overview

This paper implemented, utilized, and evaluated an automated feedback system for facilitating skill acquisition in surgical simulations through virtual reality (VR). The authors evaluated the performance of the feedback system—based on the Melbourne University temporal bone surgery simulator—through a randomized controlled trial of medical students in feedback and non-feedback groups. It was found that there was a significant improvement in the drilling performance of the feedback group, the system frequently provided timely and appropriate feedback, and that the participants found the system useful.

B. Relevant Background

VR training environments are an advantageous training tool since they provide an opportunity for repeated training in a risk-free environment. However, since expert resources are limited, previous studies have attempted to provide end-of-task summative assessments for surgical skill to overcome the need for expert supervision in VR training environments. Additionally, previous attempts of supplying real-time feedback to users have resulted in feedback on simple metrics individually, precluding meaningful and nuanced advice that human experts provide during surgical training. This study builds on previous work in developing a real-time surgical skill feedback system by introducing a system that provides feedback based on multidimensional models of surgical expertise as applied to a virtual cortical mastoidectomy procedure.

C. Methods

Participants. The study involved 24 medical students with prior knowledge of ear anatomy but no surgical experience, randomly allocated into feedback and non-feedback groups. Each participant was asked to perform the procedure twice.

Test Platform. This study utilized the University of Melbourne VR temporal bone surgery simulator. The simulator presents the trainee with 2 slightly offset images to produce the illusion of a 3D operating space, when viewed through 3D glasses. Major anatomical structures that must be identified without injury during surgery are represented in the virtual temporal bone. The user interacts with the virtual temporal bone using a haptic device resembling a surgical drill that provides three-dimensional force feedback.

Feedback System Design. A classifier was trained to recognize expert and trainee behavior from data collected from 16 expert and 11 trainee performances, including 15,455 and 20,779 "stroke" series identified in the continuous data stream output by the simulator during surgical tasks by experts and trainees respectively. Metrics characterizing stroke technique included duration, length, average speed, average acceleration, average force, straightness, median burr size, average magnification level, bone removal rate, and average distance to anatomical structures.

Feedback System Use. Upon detection of a stroke with poor surgical technique, verbal advice was relayed to participants about one of the previously mentioned metrics to best help approach the expert level. Additionally, proximity feedback was provided when the user drill tip crossed the 5 mm threshold of an anatomical structure. Surgical technique feedback was provided to the user after two repetitions of the same behavior was detected. Processing of strokes was paused for 5 seconds after feedback was presented and the same feedback would not be provided to the user within 10 seconds after it was last presented.

Data Collection. Effectiveness, accuracy, and usefulness metrics of the automated feedback system were collected. System effectiveness was measured through the percentage of strokes classified as expert within each group along with the amount of damage caused to anatomical structures. Furthermore, a post experiment evaluation was carried out by an expert otologist to assess the accuracy of the feedback system based on three error measures of false-positive classifications, incorrect feedback, and false-negative classifications. Feedback group users were interviewed to evaluate system usefulness.

Data Analysis. A confidence interval of 95% was used to test for significance ($P \leq .05$). Friedman's test, a nonparametric 2-way analysis of variance, was used for comparing performance as the data did not withstand the test for normality.

D. Results

Feedback Effectiveness. The Friedman's test adjusted for repeated trials showed that the percentage of expert strokes of the feedback group was significantly higher than that of the non-feedback group, with $\chi^2(1) = 14.450$, $P < .001$. Of all the metrics used to define stroke technique, only the bone removal rates were found to be significantly different between the 2 groups, $\chi^2(1) = 4.050$, $P = .044$, and $\chi^2(1) = 6.050$, $P = .014$, respectively, with the feedback group exhibiting elevated removal rates.

Feedback Accuracy. Feedback was provided by the system 88.6% of the time when it was necessary, and it was accurate in 84.2% of these instances.

Feedback Usefulness. Eight out of 12 participants indicated that they paid attention to the feedback, while 11 found the feedback to be helpful. However, five trainees expressed that they thought some of the feedback was contradictory or wrong. Five participants also found some feedback to be unclear.

E. Evaluation

The overall importance of this paper is its novel attempt to demonstrate how multidimensional surgical skill feedback can be integrated into procedure simulations in real-time. Furthermore, it identified significant skill metrics that can be adapted into feedback that is informative of how participants can change their behavior to approach expert performance. This paper is relevant to our project due to the skill metrics introduced, as well as its examination of methods to automatically calculate metrics of skill, limiting the need for supervised input from an instructor. Specifically, we plan to utilize stroke force and distance from anatomical structures as features in our skill classification model, as well as aim to perform real-time analysis in order to benefit surgical phase recognition.

However, we observed some limitations in this paper. While the authors noted that the feedback given by their system could be perceived as unclear or contradictory, we thought that it also shared some flaws with previous univariate studies referenced throughout the paper. Namely, we recognized that while there were significant differences in the percentage of expert strokes throughout stages of the procedure and differences in bone removal rates between experimental groups, there was really no significant difference in damage to key anatomical structures or any other performance metrics. This suggests that like the referenced papers, the researchers here had mainly succeeded in manipulating the classification of trainees in their own system, without measurably improving skill. Additionally, we realized that no details were provided on either the classification or feedback algorithms, which hinder the replicability of the study.

Based on these limitations, the next steps for this work would be to provide reproducible details on the computational methods employed, as well as to repeat the experimental procedure on individuals with more *a priori* surgical experience, such as surgical residents. This would allow for increased repricability of the research and aid in further developments being made.

IV. Conclusion

The findings and methods in both of the referenced papers assist our project to develop an objective, technical assessment of the skill of surgeons performing mastoidectomies by informing validated surgical skill metrics and implementation methods. Novel features identified for the evaluation of surgical skill level in these papers include jerk, spatiotemporal curvature, stroke force, and distance to anatomical structures. Additionally, we motivate our own work through the examination of the limitations of these papers, allowing us to prioritize the inclusion of varying levels of surgical experience and the provision of real-time skill analysis.

V. References

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

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