Analysis and Control of Temporal Biases in Surgical Skill Evaluation

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Jason David Kelly

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Timothy M. Kowalewski, PhD

May, 2020

Acknowledgements

I would like to first thank my advisor Tim Kowalewksi for his guidance throughout my time in his lab. His understanding, support, and constructive criticism have provided an immense help in completing this project. Looking back I cannot imagine completing my graduate work with any other faculty member, and I hope to work as well together in the future.

I am grateful to all the members of the Robotic Surgery Readiness study at the University of Washington and University of Minnesota including: Thomas Lendvay, Lois Meryman, Anna French, and Nicholas Heller, who were instrumental in achieving certain milestones and collecting data presented here.

Thanks to all who have been involved in the Medical Robotics and Devices Laboratory while I have been in it: Gillian McDonald, Chaitanya Awasthi, Mark Gotthelf, Mihai Duduta, Yusra Farhat Ullah, Bradley Drahos, Matthew Kubala, Ben Hamlen, Bin Fu, Mark Gilbertson, Trevor Stephens, Rebecca Smith, and several others. Whether we have worked together or not, the sense of community in our lab has made being a graduate student feel a little less like work.

I am additionally forever grateful for both of my parents, Jill and Andy, whom have always been there when I have needed them and raised me to be able to achieve new levels of success in our family. Lastly I would like to thank my wife, Katherine, who has been by my side throughout this process, and weathered all of the late nights, deadlines, failures and successes with me, helping make the journey more bearable.

Abstract

Objectively and accurately assessing the technical skill of a surgeon is critically important. The current gold standard relies on a panel of expert surgeons evaluating surgical video footage using structured survey instruments [1]. This is a prohibitively time-consuming process, thus leaving the majority of procedures unevaluated. Previous methods of evaluating skill remain prone to bias towards a surgeons' speed or task times, fueling the need to investigate the mechanisms underlying human motion in favor of techniques impervious to biases.

The research objective of this work is to investigate the effects of time and speed on the relative accuracy of both human and computational methods of measuring surgical technical skill. Human methods consist of both expert and non-expert raters (faculty surgeons and Amazon Mechanical Turk crowd workers respectively). Computational methods consist of both neurophysiologically-derived measures from other disciplines and recent model-free machine learning methods.

This research objective is pursued by the following four specific aims:

Specific Aim 1: *Determine whether surgical motion segments are directly correlated to tangential velocity prediction models, and if the result is impervious to surgeon speed.*

The objective was to test the null hypothesis that there is no relationship between the minimum jerk trajectory velocity prediction model and increases in technical skills proficiency. Prior work in human reaching suggests that adherence to the minimum jerk model should increase as technical skill increases, for proficiency of reaching motions in stroke rehabilitation. This thesis investigates whether this phenomenon holds true in surgical technical skill during simulated dry lab tasks.

Specific Aim 2: *Implement a classification algorithm which uses recorded data to classify surgical skill in a manner which is impervious to task time.*

It was hypothesized that recent machine learning algorithms which exploit temporal duration, used on kinematic data from dry lab laparoscopic training tasks, would increase the performance of the current state-of-the-art computational methods of classifying surgical skill.

Specific Aim 3: *Determine how human ratings of surgical tasks are affected by video playback speed and duration.*

It was hypothesized that the perceived skill of a surgeon followed a unimodal function, in which human raters would experience an increase in perceived surgical skill as the speed of a surgical task video approaches the function's maximum, then immediately decreasing once being aware of the video's playback manipulation.

Specific Aim 4: *Measure the effect that pre-operative warm-up using validated virtual-reality simulator tasks has on practicing surgeons in real robotic surgeries as measured by the most accurate, least-biased methods detailed in the previous specific aims and prior art.*

This tested the hypothesis that pre-operative warm-up results in a measurable improvement in surgical technical skill among practicing surgeons (no novices) using surgical robots from live patients.

This research concluded firstly that neurophsyiologically-derived models of skill, specifically the minimum jerk model, do not necessarily extend to surgical settings. Surprisingly this research found the opposite, that surgeon experts exhibit movements which deviate further from the minimum jerk model. Second, a classification algorithm was created, using a bidirectional long short-term memory network which controls for task time, and is capable of classifying experts and novices with over 95% accuracy for tasks most resembling real surgery. This research brought about questions of label noise and accuracy, and emphasizes the importance of properly labeled data for machine learning algorithms. It was found that humans appear to have a speed bias in rating surgeons both for laparoscopic surgical training tasks as well as real robotic surgery procedures. Unexpectedly, this effect was more substantial for more expert performances and negligible for novice performers. Counter to our original hypothesis, expectations derived from biological motion models – that skill discrimination capability would unimodally decrease when video playback was obviously artificially sped up – were not met. Observer ability to discriminate skill continues well after people are cognizant of a video being played at quicker speeds and no discernible difference between biological-motion-relevant question groups (e.g. motion fluidity) and other questions appeared. Finally, a new dataset of robotic surgeries was introduced, with 343 videos of robotic surgeries including tooltip kinematic data. Evidence obtained from motion metrics, crowd ratings, and faculty surgeon ratings suggest that no measurable warm-up effect was present in our population of 41 practicing surgeons; no evidence supported the use of warm-up.

Contents

Acknowledgements i

Abstract ii

List of Tables ix

List of Figures xi

1 Introduction 1

2 Background 4

2.1 Project Description 5

2.1.1 Motivation and Objectives 5

2.1.2 Crowd Sourcing 6

2.1.2.1 Likert Scale Scoring 8

2.1.3 Hand Movement Research 12

2.1.3.1 Measures of Movement Smoothness 12

2.1.3.1.1 Submovements 12

2.1.3.1.2 Spectral Arc Length 13

2.1.3.1.3 Minimum Jerk 16

2.1.3.2 Objective 18

2.1.4 Computational Skill Evaluation 18

2.1.4.1 Neural Networks 18

2.1.4.2 Recurrent Neural Networks 21

2.1.4.3 Long Short-Term Memory 21

2.1.4.4 Automated Performance Metrics . . . 23
2.1.4.5 Video Motion Data Analysis . . . 23
2.1.4.6 Lack of Quality Data . . . 26
2.1.4.7 Objective . . . 28
2.1.5 Biological Motion Perception . . . 28
2.1.5.1 Visual Stimuli Research . . . 29
2.1.5.2 Response to Abnormal Dynamics . . . 30
2.1.5.3 Objective . . . 30
2.1.6 Datasets to be used in this work . . . 32
2.1.6.1 Basic Laparoscopic Urologic Study . . . 32
2.1.6.2 Robotic Surgery Readiness Study . . . 32
2.1.7 Linear Mixed Effects Modeling . . . 33

3 Neurophysiological Models **36**
3.1 Testing the Relationship of Minimum Jerk Model to Surgical Skill . . . 36
3.1.1 Methods . . . 36
3.1.2 Results . . . 37

4 Computational Skill Evaluation **44**
4.1 Bidirectional Long Short-Term Memory for Surgical Skill Classification of Temporally Segmented Tasks [69] . . . 45
4.1.1 Abstract . . . 45
4.1.2 Introduction . . . 46
4.1.3 Methods . . . 47
4.1.3.1 Dataset . . . 47
4.1.3.2 Skill Classification from Temporal Segmentation . . . 49
4.1.3.3 Data Partitioning . . . 49
4.1.3.4 Window Parameter Selection . . . 49
4.1.3.5 LSTM Parameters and Architecture . . . 50
4.1.3.6 Crowd Reassessment . . . 51
4.1.4 Results . . . 52
4.1.4.1 Skill Classification from Temporal Segmentation . . . 52
4.1.4.2 Crowd Reassessment . . . 57

4.1.5 Conclusion . . . 60
4.2 Classification Using Video Data . . . 62

5 Speed Perception **65**
5.1 The Effect of Video Playback Speed on Surgeon Technical Skill Perception [73] . . . 66
5.1.1 Abstract . . . 66
5.1.2 Introduction . . . 67
5.1.2.1 Biological Motion Perception . . . 68
5.1.2.2 Objective Metrics . . . 68
5.1.3 Methods . . . 71
5.1.3.1 Dataset . . . 71
5.1.3.2 Experiment 1: Technical Skill Perception . . . 72
5.1.3.3 Experiment 2: Speed Perception . . . 73
5.1.3.4 APM Validation at Each Playback Speed . . . 73
5.1.4 Results . . . 74
5.1.4.1 Technical Skill Perception . . . 74
5.1.4.2 Playback Speed Perception . . . 76
5.1.5 Conclusion . . . 77
5.2 The Effect of Video Playback Speed on Perception of Technical Skill in Robotic Surgery [81] . . . 83
5.2.1 Abstract . . . 83
5.2.2 Introduction . . . 84
5.2.2.1 Changes in Perception due to Speed . . . 86
5.2.2.2 Technical Skill Ranges . . . 87
5.2.3 Methods . . . 88
5.2.3.1 Dataset . . . 88
5.2.3.2 Technical Skills Perception at Different Playback Speeds 89
5.2.3.3 Sub-Task Level Skill Labeling . . . 90
5.2.4 Results . . . 90
5.2.5 Data Demographics . . . 90
5.2.5.1 Technical Skill Perception . . . 92

5.2.5.2 Sub-Task Level Labeling 93
5.2.6 Conclusion 94

6 Robotic Surgery Readiness Study 96
6.1 Robot-Assisted Surgery Readiness: Virtual Reality Warm-Up Prior to Robot-Assisted Surgery [93] 96
6.1.1 Abstract 96
6.1.2 Introduction 98
6.1.3 Methods 101
6.1.3.1 Aim 1 101
6.1.3.1.1 Subject Recruitment 101
6.1.3.1.2 dVSS Modules 102
6.1.3.1.3 Proficiency Testing 102
6.1.3.1.4 Video Review 104
6.1.3.1.5 Statistical Models 104
6.1.3.2 Aim 2 105
6.1.3.2.1 Subject Recruitment 105
6.1.3.2.2 Randomization and Warm-Up Protocol 105
6.1.3.2.3 OR Data Capture 105
6.1.3.2.4 Data Aggregation and Manipulation 106
6.1.3.2.5 Video Segmentation 106
6.1.3.2.6 Objective Metric Calculation 106
6.1.3.2.7 Video Review 107
6.1.4 Results 107
6.1.4.1 Aim 1 107
6.1.4.2 Aim 2 109
6.1.4.2.1 GEARS Review Outcome 109
6.1.4.2.2 Objective Metric Outcomes 113
6.1.5 Discussion 116
6.1.6 Conclusions 117
6.2 Temporal Variability in Surgical Technical Skill Evaluation [117] 118
6.2.1 Abstract 118

6.2.2 Introduction . . . 119
6.2.3 Methods . . . 121
6.2.3.1 Dataset . . . 121
6.2.3.2 Crowd Evaluation . . . 122
6.2.3.3 Statistical Analysis . . . 123
6.2.4 Results . . . 124
6.2.5 Conclusion . . . 126

7 Conclusion and Discussion **128**
7.1 Contributions . . . 129
7.2 Limitations . . . 131
7.2.0.0.1 Speed Perception Studies . . . 131
7.2.0.0.2 Long Short-Term Memory Study . . . 131
7.2.0.0.3 Robotic Surgery Readiness Study . . . 131
7.2.0.0.4 Rating Variability Study . . . 131
7.3 Future of Surgical Skill Research . . . 132

References **133**

Appendix A. Robotic Surgery Readiness Supplemental Figures **149**
A.0.1 Kinematic Metrics . . . 153
A.0.2 Event Metrics . . . 153

List of Tables

2.1 Likert-scale technical skill perception questionnaire for video-based evaluation of a surgical task, from four domains of the GOALS assessment metric [24]. 10
2.2 Likert-scale technical skill perception questionnaire, from the five of the 6 domains in the GEARS assessment metric. Autonomy is left out, as no audio or video of the operating room is provided in our datasets. 11
2.3 Summary of popular skill evaluation studies, and lack of high volume data. 27
3.1 ANOVA test results for the mixed effects model on the minimum jerk viability as a predictor variable for task score value. 40
4.1 Main hyperparameters tested during model evaluation in Section 3.1, with most optimal results in bold. 53
4.2 Accuracy results computed by the bidirectional LSTM for each BLUS task. 55
4.3 Correlation coefficients for the relationship between GOALS scores and predicted scores from the LSTM for every performance from each of the four BLUS tasks, including intermediate level performers, from Section 3.1. 55
5.1 Likert-scale speed perception questionnaire for a manually sped-up video. 68
5.2 Results from the linear mixed effects model testing speed and time spent on evaluation 90
5.3 The types of robotic surgery performed and demographics of the surgeons from the 56 videos used. 91
6.1 The five groups of warm-up modules used for aim 1. Each group consisted of different combinations of warm-up modules shown to help elevate skill. 104
6.2 Results from linear mixed model tested with kinematic and event outcomes. 115

6.3 Results from Friedman test with kinematic and event outcomes. 115
6.4 Correlation values for each of the metrics measured with GEARS scores. 115
6.5 Summary information for 12 videos used from the RSR study. 122
A.1 Recorded event measures for the RSR study. 153

List of Figures

2.1 Relationship between speed of task completion and crowd assessed score. Surgeons with a longer task time on average receive lower scores, as rated by crowds. 6

2.2 Original normal curve found through Galton's crowd-sourcing experiments with ox weights [17]. The mean of these data converge to the correct weight of the ox, within one pound. 7

2.3 Previous research found a high correlation between faculty surgeon review and non-expert crowd review of surgical technical skill [21] 8

2.4 Typical movements(bold lines) and their extracted submovements(fine lines) from first and last days of a therapy session for a stroke patient [11]. 12

2.5 The effect of the SAL metric on varying lengths of movements [28]. Longer movements provide a lower SAL measurement. 15

2.6 The effect of temporal movement scaling on SAL vs. SPARC [28]. SPARC is relatively impervious to movement scaling. 16

2.7 Predicted(solid lines) and measured(dashed lines) hand trajectories overlaid [3]. The minimum jerk model provides accurate predictions for the movements of recovering stroke patients. 17

2.8 An example of the architecture of a neural network [34]. Inputs are sent through layers in which activation functions and array manipulations are used to arrive at a single output signal. 20

2.9 An example of the architecture of a long short-term memory network. A series of gates are used to create a sense of context for past mistakes and successes. 22

2.10 The detected SIFT points from a dry lab laparoscopic surgical task image (computed using MATLAB computer vision toolbox). 24
2.11 Comparison of corner detection boundaries for Harris and Shi-Tomasi detection algorithms. Shi-Tomasi has a more rigid set of constraints, and will detect more corners because of these boundaries. 25
2.12 Top: Discriminated motion counts corresponding to each class of STIPs. Bottom: Frame kernel matrices example, displaying the relationship between various segments of a video [55]. 27
2.13 Example of a point light walker. Each of the dots represent a limb or joint of the human body. 29
2.14 Illustration of a PLW representing an animal (right) and the change in perceived size as the gait increases (left). This suggests a link between speed and biological motion perception [66]. 30
2.15 Frames from a (left to right) suturing, cutting, and peg transfer task in the BLUS dataset. 32
2.16 Range of crowd-assessed surgical skill evaluation mean scores for each of the BLUS tasks. On average clipping tasks receive higher scores than, for example, suturing tasks. 33
2.17 Frame from a video in the RSR data set, recorded using a daVinci surgical robot (Intuitive Surgical (Sunnyvale, CA). Typical robotic surgeries recorded were prostatectomy, hysterectomy, and partial nephrectomy, with all procedures being recorded in Washington state. 34
3.1 All Peg Transfer task segments overlaid to show the general shape of peg transfer task grasp segments. One single line represents one single movement by a surgeon. Movements were defined as starting and ending when the hand reached nonzero speed. Time is normalized by the duration of each segment and velocity is scaled by the peak velocity magnitude of that segment. 38
3.2 Example of segmented grasps, from the first performance of the peg transfer tasks only. Showing the data in this manner allows seeing the variability that exists throughout the different movements in a surgical task performance. 39

3.3 Unimodal movements from peg transfer tasks fitted to minimum jerk movements described in Chapter 2.1.3.1.3. 40
3.4 Fixed effects model of observed vs. predicted skill, for all BLUS tasks combined, using minimum jerk model, subject and task identity, in which the fixed effects accounts for a constant intercept for all of these factors. The fixed effects model behaves quite poorly. 41
3.5 Linear Mixed effects model's predicted vs. observed skill using minimum jerk model fit error, task, and subject indentity, in which this model accounts for varying intercept values for each task, and subject. This model is able to predict outcomes much better than the fixed effects model, with a correlation between actual and predicted skill of 0.826. . . 42
3.6 LME predictions grouped by task, with colorbar illustrating total time for task completion. There is a strong correlation between observed task scores and predicted scores, with slower performances usually receiving poorer of both. 43
4.1 Example left hand tool acceleration time series with the sliding time window extraction method applied to both a novice series and an expert series. A novice performance is indexed with a window overlap of 105 time indices, and an expert performance is indexed with a window overlap of 50 time indices. 51
4.2 Process of computing final suturing performance prediction. All ten candidate predictions of a performance are averaged at each index to compute the final prediction. 52
4.3 The LSTM's prediction for all performances from each of the tasks in the BLUS dataset. The LSTM was only trained on experts and novices (summary score labels with cross validation), from Section 3.1. 54
4.4 The LSTM predictions trained on crowd scores, for performances of suturing tasks as rated by faculty surgeons, which has a Spearman correlation of 0.89, from Section 3.1. 56
4.5 The perceived score of the LSTM on the left, with a video on the right of the same task. 57

4.6 Reassessment of misclassified suturing and peg transfer performances suggest crowds agree more with LSTM than previous crowd ratings, from Section 3.2. 59
4.7 A bar plot of the magnitude of attenuation of the crowds to agree more with the LSTM on score reassessment, than with the original scores obtained. 60
4.8 The frames of a novice video with the amount of points belonging to each motion class being shown. 62
4.9 The frames of an expert video with the amount of points belonging to each motion class being shown. 63
4.10 Frame kernel matrix for a portion of a time series, to visualize similarity between frames. 63
5.1 All mean crowd evaluations from each novice and expert peg transfer task at various video playback speeds. (Each solid marker indicates N = 40). 74
5.2 The efficiency subdomain as compared to the average of the other GOALS subdomains, for experts and novices. 75
5.3 Objective technical skill metrics for the obvious experts and novices across various speeds. 76
5.4 Difference in mean between skill levels from the highest performing APM and from crowd scores, with a 95% CI in the shaded regions. 77
5.5 Comparison of Novice and Expert Speed Perception at each playback speed. (There were $N = 40$ *unique human evaluations/skill level/playback speed*. The error bars represent 95% confidence intervals.) 78
5.6 Speed Perception compared with technical skill perceived at various speeds. 79
5.7 Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is $N = 40$, with 3600 total evaluations. 81
5.8 Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is $N = 40$, with 3600 total evaluations. 82
5.9 Frame from a video collected for this study, recorded using a daVinci surgical robot, Intuitive Surgical (Sunnyvale, CA). 85

5.10 Illustration of research results (left) which found the frequency of gait in simulated animals (right) affects the estimated size [**?**]. 86
5.11 All mean crowd evaluations (bold lines) from each proficient and expert surgeon at various video playback speeds. Each single surgeons video (semi-transparent colors) indicates ratings from N = 40 turkers. 91
5.12 The efficiency subdomain as compared to the mean of the other GEARS subdomains, for expert and proficient surgeons. 92
5.13 GEARS scores given to the entire 15 minute video of a performance, compared to only the first minute of the same video. 93
6.1 The popularity in robotic surgery has experienced a significant increase in recent years. 99
6.2 The da Vinci Skill Simulator (dVSS) used for surgical simulation modules, created by Mimic Technologies. 100
6.3 The four dVSS warm-up modules used for aim 1 of the RSR study including Ring and Rail 2, Match Board 3, Suture Sponge 3, and Running Suture (left to right). 103
6.4 The randomization setup for the 5 warm-up groups in aim 1 of the RSR study. 103
6.5 Enrollment demographics for participants in aim 1 of the RSR study. . . 108
6.6 The task time (TT) for the five different warm-up module groups, throughout the six completed sessions. 110
6.7 The GEARS scores for the five different warm-up module groups, throughout the six completed sessions. 111
6.8 Enrollment demographics for participants in aim 2 of the RSR study. . . 112
6.9 The correlation of the GEARS scores from faculty surgeons and C-SATS scores was weak, with a value of 0.1714, 95% CI from -0.1287 to 0.4426. 113
6.10 The effect of Si and Xi da Vinci platforms for surgeons who performed and did not perform a warm-up module before surgery. No significant difference exists between these four groups. 114
6.11 The effect of performing a warm-up module for each each surgeon in the RSR study. With a clear normal distribution, no significant effect was measured. 116

6.12 The attention question inserted to the GEARS questionnaire, for quality assurance purposes. Note that skimming the instructions likely results in incorrect answers. 123
6.13 All mean crowd evaluations from each proficient (red) and expert (blue) surgeon at each minute of the performance. Surgeons previously randomized to the control (solid lines) group normally started surgery without any intervention or pre-operative warm-up used. Warm-up (dashed lines) group surgeons reviewed a virtual reality warm-up module prior to surgery. (The warm-up hypothesis from the original randomized study is not being tested or evaluated in this research, only temporal variation in ratings is. 124
6.14 Comparison of previously-obtained scores for 15 minute video (95% CI) and average score for evaluation of 1 minute segment. (PW = Proficient/Warmup; PC = Proficient/Control; EW = Expert/Warmup; EC = Expert/Control) 125
6.15 Standard deviation from the LME model for random effects from each individual crowd rater, compared with the fixed effect of time. The rater effect is clearly larger, although time is still significant in predicting outcomes. 126
A.1 Video of an Si daVinci surgery, after editing the video to the correct times and placing black boxes over the identifying pieces of information. . . . 150
A.2 3-D Plot of the kinematic tooltips from a robotic surgery which corresponds with video edited at the same time range. USM 1-3 refers to the Universal Serial Manipulators, the surgical arms. 150
A.3 A view of the website faculty surgeons would see when visiting to evaluate robotic surgeon technical skill. 151
A.4 Cumulative positions of several combined tasks from the RSR study, in both 3-D coordinates, as well as a histogram for frequency of position in each axis. 152

Chapter 1

Introduction

This thesis aims to discuss previous and current work in computational surgery, and more specifically, surgical skill evaluation. Skill evaluation is a topic becoming popular in several research labs, but it is apparent that accurately evaluating technical skill is a multi-faceted problem, with several specific areas of research which should be investigated. Moreover, previous research appears to reward time of task completion, perhaps unfairly, to surgeons. The work presented in this thesis aims to both summarize previous work as well as test the existence of this bias in surgical settings and create techniques which control for these biases.

The research from this thesis begins by examining hand movements models which were derived from neurophysiology research involving stroke patients (Chapter 3). It was found that the form of hand movements of recovering stroke patients most closely agreed with the minimum jerk velocity profile. To test these models and control for the speed of performances, individual point-to-point movements from surgical tasks were extracted and normalized to use with statistical models. The results show a surgeon is surprisingly more likely to be novice-like if they exhibit movements which correspond to the minimum jerk model, contrary to what prior research would suggest.

Summary metrics of performances can be biased to give quicker surgeons better metric scores due to the way they are calculated. In Chapter 4 machine learning models were used in an attempt to classify the skill of a surgeon, in a way which does not bias task time. Multidimensional kinematic data was used to train a model, in which the data was segmented into small multi-second chunks. Training the model in this manner

allowed the model to learn aspects of movements without being biased to reward faster or slower surgeons. This was able to classify surgeons into skill levels with over 95% accuracy for suturing tasks which most closely resemble surgical procedures. However, reassessing misclassified surgeons led to questions about label noise, and reiterates the need for accurate labeling of surgical data.

The third aim was then addressed by examining the effect of video playback speed on surgical skill perception (Chapter 5). The goal was to learn whether humans perceive technical surgical skill differently when viewing video at quicker playback speeds, and if this continues once they are aware of the video being sped up. When working with laparoscopic simulation tasks it was found that skill was more easily discriminated when a video was manipulated to play faster. Additionally, this phenomenon extended past the point at which users were aware of the video being sped up, rejecting the null hypothesis that the perceived increase in skill was reliant on being unaware of the playback speed being changed. In an effort to create a more statistically powerful finding, this study was repeated for a larger dataset of robotic surgery procedures, which also found increased perceived skill levels at quicker playback speeds.

Finally, the fourth aim was addressed in Chapter 6, during which a new dataset of robotic surgery procedures was introduced. The results from this research, as measured by motion metrics, crowd scores, and faculty ratings, concluded that participating in a virtual reality warm-up module prior to robotic surgery did not significantly elevate the technical proficiency of the surgeons involved in this study. Nevertheless, this dataset will allow new and innovative techniques to be used in the future in pursuit of solving surgical skill evaluation challenges. This chapter also reviews work examining the temporal variability in skill evaluation, finding that for longer duration surgical procedures, summary metrics may hinder in properly conveying the overall skill of a surgeon. Measures which report the variability of surgical skill may be required.

This thesis will conclude with a chapter summarizing the main findings of this research, discussing limitations of the work presented, and discussing future work in the field of surgical skill evaluation (Chapter 7).

- **Chapter 1: Introduction**
- **Chapter 2: Background** - Briefly presents the history, related work, and need for improvement in the area of computational surgical skill evaluation.
- **Chapter 3: Neurophysiological Models** - Describes the process of evaluating the relationship between the minimum jerk model and surgical skill.
- **Chapter 4: Computational Surgical Skill Evaluation** - Provides an introduction to a machine learning algorithm which can accurately evaluate the technical skill of a surgeon.
- **Chapter 5: Speed Perception** - Provides an analysis and explanation for how people could be biased in evaluating surgical skills based on the perceived speed of a surgeon.
- **Chapter 6: Robotic Surgery Readiness Study** - Presents the need for the robotic surgery readiness study, explains the methods of gathering data, and the results of the study, in addition to studying how human perception of skill can vary in longer duration recordings.
- **Chapter 7: Discussion** - This chapter concludes the thesis work and provides guidance for how to proceed and improve upon this work in the future.

Chapter 2

Background

Large 'crowd-sourced' groups of non-experts are surprisingly accurate at evaluating surgical tasks [1], though it remains unclear as to 'how' or 'why'. However, as shown in [2], human perception can be affected by various physical features such as size or speed. Due to this, work must be done to create an unbiased metric for viewing complex human motions using approaches including algorithms impervious to bias. Previous research has shown that optimized velocity trajectories of point-to-point movements can be predicted using methods such as minimizing the jerk of the tangential velocity of a movement [3]. However, surgeries involve very complex movements, and the mapping of these trajectories to a performer's technical skill is not yet clear.

The objective of this research is to elucidate the mechanisms that underlay human perception of surgical skill, computational models, and velocity trajectory movements which correlate to surgeon technical performance, and accurately measure the perception of skill as a function of video playback speed.

2.1 Project Description

2.1.1 Motivation and Objectives

The prevalence of surgical errors in the U.S. makes it essential to develop algorithms capable of learning how surgical skill can be classified [4], [5]. Part of this effort is learning what features of a procedure can be objectively determined to be expert-like. In the past, hidden markov models have been used to assess surgeon skill, but these are not incredibly accurate [6–8]. While they perform well with small amounts of data, they are not able to become increasingly accurate as more data is available. Compared to newer machine learning techniques, hidden markov models are relatively simple and will not be able to extract as complex information as other more recent techniques [9]. Much success in other academic areas has been the result of variants of recurrent neural networks, which can arm a model with a sense of memory to remember previous events. Using a model such as this with kinematic features from tooltips would be an immensely valuable way to classify surgical skill.

There is a plethora of previous research regarding skill, claiming certain tangential velocity hand movement models are related to skill of a performer, but much of this only relates to simple point-to-point movements [3, 10–12]. Surgery involves complex, but also repetitive movements. There are many factors at play which influence a surgeon's relative skill. The single metric which most closely correlates to the skill of a surgeon is using the time of task completion, or task time [13]. This relationship can be seen in Figure 2.1 for the BLUS peg transfer task. This is not perfect by any means though, and is just one of the factors at play when evaluating a surgeon's skill. There must be more in depth research done to evaluate this phenomenon, and learn whether phenomenal surgeons operating slowly can be perceived as experts or if crowds are rating fast surgeons well due to the quick movements they are seeing.

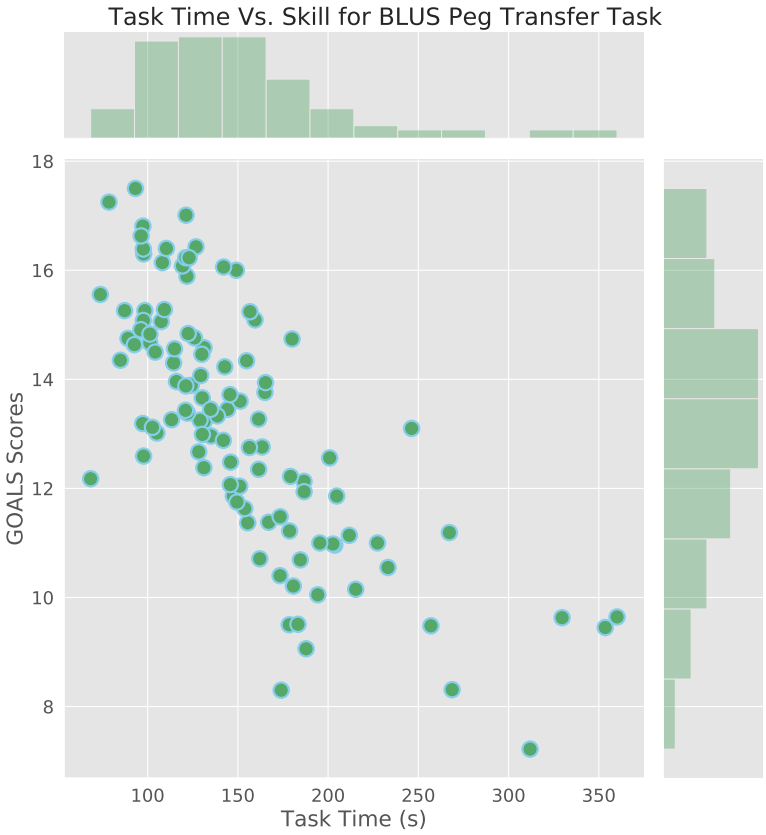
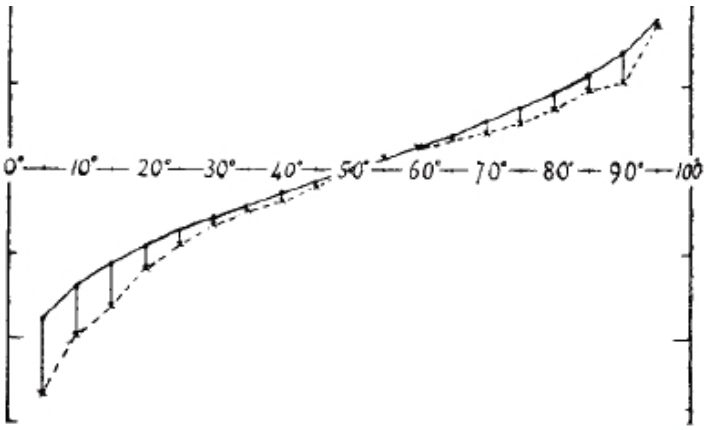


Figure 2.1: Relationship between speed of task completion and crowd assessed score. Surgeons with a longer task time on average receive lower scores, as rated by crowds.

2.1.2 Crowd Sourcing

Previous research has shown that when a large population of diverse people are asked certain questions, the truth emerges from the responses as the population size grows [14]. [15] defines crowdsourcing as, "*the act of outsourcing tasks originally performed inside an organization, or assigned externally in form of a business relationship, to an undefinable large, heterogeneous mass of potential actors*". Essentially, collective knowledge from a crowd can more easily reflect the true answer of specific types of questions, than can any one person. This was first discovered by statistician Sir Francis



The continuous line is the normal curve with p.e.=37.
The broken line is drawn from the observations.
The lines connecting them show the differences between the observed and the normal.

Figure 2.2: Original normal curve found through Galton's crowd-sourcing experiments with ox weights [17]. The mean of these data converge to the correct weight of the ox, within one pound.

Galton, the father of "eugenics". Galton famously came across a county fair competition in 1906, in which visitors were asked to guess the weight of an ox. He later used the responses to run some statistics. To his surprise, the mean of all the guesses resulted in 1197 pounds, the actual weight being 1198 pounds (Fig. 2.2) [16, 17]. This finding was later elborated on, finding that the crowds needed to be both large and diverse, or this would not work accurately. There are debates about how many workers must be used to attain a satisfactorily 'true' answer, but generally, higher populations yield more accurate results [18].

These methods have also been used in the past to allow rating skill of a surgical task or study [19]. The current gold standard for rating surgeons is a faculty panel of surgeons which rate surgeries one at a time. This is very time-consuming and suboptimal, as faculty surgeons could spend their time with more valuable tasks. Crowdsourcing workers to rate surgical tasks has proved to be a valid method of evaluation with high correlation to the gold standard (Fig. 2.3) [20, 21].

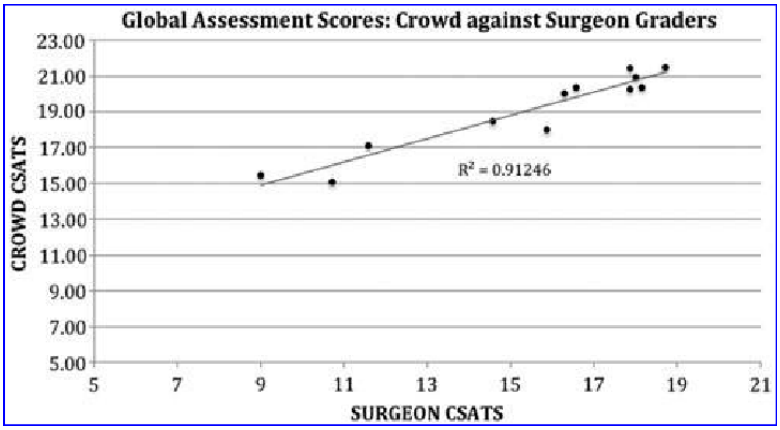


Figure 2.3: Previous research found a high correlation between faculty surgeon review and non-expert crowd review of surgical technical skill [21]

There are several platforms currently used for crowdsourcing tasks, such as Idea Bounty, Innocentive, CrowdSpring, and Amazon Mechanical Turk. Amazon's Mechanical Turk has become popular for businesses and companies to be able to request workers to complete "Human Intelligence Tasks", or HITs, for some amount of compensation. Very simple tasks such as completing surveys or watching videos, could be assigned. Once workers gain experience, and gain reliability, users can be offered more compensation to complete tasks, which filter out workers that may not put in a great amount of effort or focus. Mechanical Turk has been used previously in surgical crowdsourcing studies, finding that crowds give a 100% matching pass/fail rate of what faculty surgeons would rate when reviewing video of surgeons [1], [19, 22, 23].

2.1.2.1 Likert Scale Scoring

A Likert scale is a type of scale in which the user has a choice of several pre-defined options for how to answer. This could be a question in which the user answers with numbers from 1-10, or even answers such as "Likely", "Not at All Likely", "Very Likely", and so on. When both faculty surgeons and non-expert crowd workers evaluate surgical

tasks, they are most likely using a Likert scale to assess the surgeon. Two assessment methods will particularly be discussed several times throughout this research, the Global Evaluative Assessment of Robotic Skills (GEARS) and the Global Operative Assessment of Laparoscopic Skills (GOALS). These consist of five and four subdomain questions, respectively, which ask the rater to evaluate various aspects of the surgeon's skill, such as depth perception, tissue handling, or robotic control, on a scale of 1-5. Scores for each subdomain are totaled for a range of cumulative scores for GOALS being between 4-20 and for GEARS being 5-25, with higher scores inferring a more expert surgeon. Both scales are shown in Table 2.2 and Table 2.1.

Table 2.1: Likert-scale technical skill perception questionnaire for video-based evaluation of a surgical task, from four domains of the GOALS assessment metric [24].

Score	**Depth Perception**
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	
(3)	Some overshooting or missing of target, but quick to correct
(4)	
(5)	Accurately directs instruments in the correct plane to target
Score	**Bimanual Dexterity**
(1)	Uses only one hand, ignores non-dominant hand, poor coordination between hands
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	
(5)	Expertly uses both hands in a complementary manner to provide optimal exposure
Score	**Efficiency**
(1)	Uncertain, inefficient efforts; many tentative movements; constantly changing focus or persisting without progress
(2)	
(3)	Slow, but planned movements are reasonably organized
(4)	
(5)	Confident, efficient and safe conduct, maintains focus on task until it is better performed by way of an alternative approach
Score	**Tissue Handling**
(1)	Rough movements, tears tissue, injures adjacent structures, poor grasper control, grasper frequently slips
(2)	
(3)	Handles tissue reasonably well, minor trauma to adjacent tissue(i.e. occasional unnecessary bleeding or slipping of the grasper)
(4)	
(5)	Handles tissues well, applies appropriate traction, negligible injury to adjacent structures

Table 2.2: Likert-scale technical skill perception questionnaire, from the five of the 6 domains in the GEARS assessment metric. Autonomy is left out, as no audio or video of the operating room is provided in our datasets.

Score	**Depth Perception**
(1)	Constantly overshoots target, wide swings, slow to correct
(2)	
(3)	Some overshooting or missing of target, but quick to correct
(4)	
(5)	Accurately directs instruments in the correct plane to target
Score	**Bimanual Dexterity**
(1)	Uses only one hand, ignores non-dominant hand, poor coordination
(2)	
(3)	Uses both hands, but does not optimize interaction between hands
(4)	
(5)	Expertly uses both hands in a complementary way to provide optimal exposure
Score	**Efficiency**
(1)	Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress
(2)	
(3)	Slow, but planned movements are reasonably organized
(4)	
(5)	Confident, efficient and safe conduct, maintains focus on task, fluid progression
Score	**Force Sensitivity**
(1)	Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage
(2)	
(3)	Handles tissue reasonably well, minor trauma to adjacent tissue, rare suture breakage
(4)	
(5)	Applies appropriate tension, negligible injury to adjacent structures, no suture breakage
Score	**Robotic Control**
(1)	Consistently does not optimize view, hand position, or repeated collisions even with guidance
(2)	
(3)	View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant.
(4)	
(5)	Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant.

2.1.3 Hand Movement Research

Extensive point-to-point hand movement research has been used to create models which correlate to skill in non-surgical settings [10, 11, 25]. Notably, the minimum jerk model has been used in the past with studies of stroke patients, to define the skill involved in a movement, for various weeks after recovery [3]. It was hypothesized that more skilled movements were smoother. It is known that when hand movements occur, there are submovements present which appear in the form of several individual movements occurring in sequence [10]. Potentially, smoother movements lead to more 'expert' behavior or performance.

2.1.3.1 Measures of Movement Smoothness

2.1.3.1.1 Submovements It has been observed that early movements in recovering stroke patients exhibit movement 'fragments', known as submovements, shown in Figure 2.4 [11]. These submovements have been used in several studies, to demonstrate different metrics for quantifying skill level in movement tasks. The main area of interest is a movement's smoothness, or the continuity and non-intermittency of a movement. This smoothness increases with learning and experience [26].

Figure 2.4: Typical movements(bold lines) and their extracted submovements(fine lines) from first and last days of a therapy session for a stroke patient [11].

Submovements may be used as a measure of how smooth a movement may be. More

submovements that occur between two points should lead to lower overall smoothness and less ability to perform the task. Decomposing a movement into its' submovements is a nonlinear global optimization problem, to determine the correct parameters suitable to fit the submovements to smaller movements. An example of this is shown in Figure 2.4. A 'scattershot' algorithm has been introduced that finds the globally optimal submovements' composition, by calculating the probability of finding the globally best fit [11].

This scattershot algorithm uses support-bounded log normal(LGNB) curves, which can be produced with

$$B(t) = \frac{D(T_1 - T_0)}{\sigma\sqrt{2\pi}(t - T_0)(T_1 - t)} \exp\left\{\left(\frac{-1}{2\sigma^2}\right)\left[ln\left(\frac{t - T_0}{T_1 - t}\right) - \mu\right]^2\right\} \tag{2.1}$$

from $T_0 \leq t \leq T_1$, where D is the displacement resulting from the movement, T_0 is the movement start time, T_1 is the end time, μ controls skewness, or assymetry, and σ determines kurtosis, or fatness, of the curve. This function was created by Palomondon, and describes the movements as originating from the sequential action of a set of velocity generators working in a cascaded fashion. Velocity profiles can be described by LGNB curves which are asymmetric and continuous [25].

2.1.3.1.2 Spectral Arc Length Measuring smoothness of a movement can be done with several different metrics, such as the number of peaks in the velocity profile, the minimum squared jerk profile, and the speed profile. The most recent and successful metric is called the spectral arc length (SAL) [27]. SAL uses the Fourier magnitude spectrum of the speed profile, and measures the arc length of the movement's normalized

Fourier magnitude spectrum. This is defined as:

$$\eta_{sal} \triangleq -\int_0^{\omega_c} \sqrt{\left(\frac{1}{\omega_c}\right)^2 + \left(\frac{d\hat{V}(\omega)}{d\omega}\right)^2 d\omega}, \tag{2.2}$$

$$V(\omega) \triangleq \frac{V(\omega)}{V(0)}, \tag{2.3}$$

where V(ω) is the Fourier magnitude spectrum of $v(t)$, and $[0, \omega_c]$ is the frequency band occupied by the given movement. ω_c is set to 40π rad/s, which covers normal human movement. $\hat{V}(\omega)$ is the amplitude normalized magnitude spectrum [27]. When SAL was used to test the smoothness of recovering stroke patients, and compared with other popular smoothness metrics, SAL was the only metric used in which the smoothness improved as the patients progressed [3].

One downside of SAL is that it computes a higher smoothness for tasks that occur more quickly, instead of just analyzing the movement itself, shown in Figure 2.5. An improvement was recently made to SAL, called SPARC(SPectral ARC length), which is independent of temporal movement scaling, as shown in Figure 2.6 [28].

It should be noted that movement smoothness is highly task dependent, and should only be compared among similar tasks, with similar constraints. The kinematic data used to measure smoothness should also be a derivative of a position metric, like velocity or jerk. This is because the smoothness measures will highlight whether intermittency occurs in the movement and ignore features that are not changing with time.

Smoothness measures can have large differences depending on whether or not the measure is being recorded over a discrete movement, in which the movement goes from one finite point to another. If the movement is instead rhythmic, in which, for example,

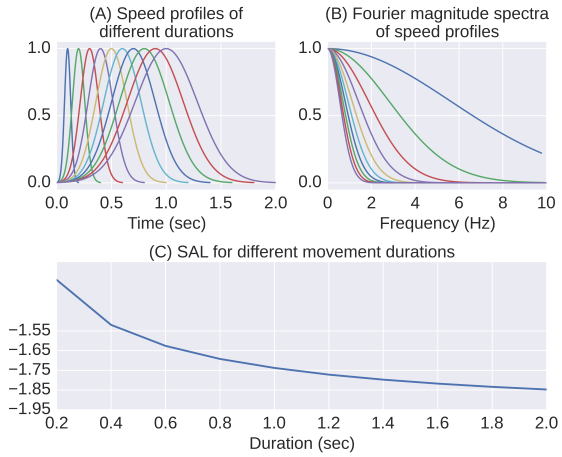


Figure 2.5: The effect of the SAL metric on varying lengths of movements [28]. Longer movements provide a lower SAL measurement.

a person's hand is moving back and forth in the same relative areas, repeatedly, a different method of measurement must be computed. If x(t) is the movement, then it can be represented as a concatenation of individual movements,

$$x(t) = \sum_{i=1}^{N} x^i(t); \quad x^i(t) = x(t)\Pi_i(t);$$
$$\prod_i(t) = \begin{cases} 1, & \text{if } t_i \leq t < t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

in which $x^i(t)$ is the ith movement, multiplied by the rectangular window $\prod_i(t)$ of the movement. t_i is the start time of movement, t_{i+1} is the end of the movement. If a perfectly rhythmic movement is used, t_i can be chosen such that the concatenation only captures certain portions of a movement.

The smoothness of each movement can then be found independently. In order to do

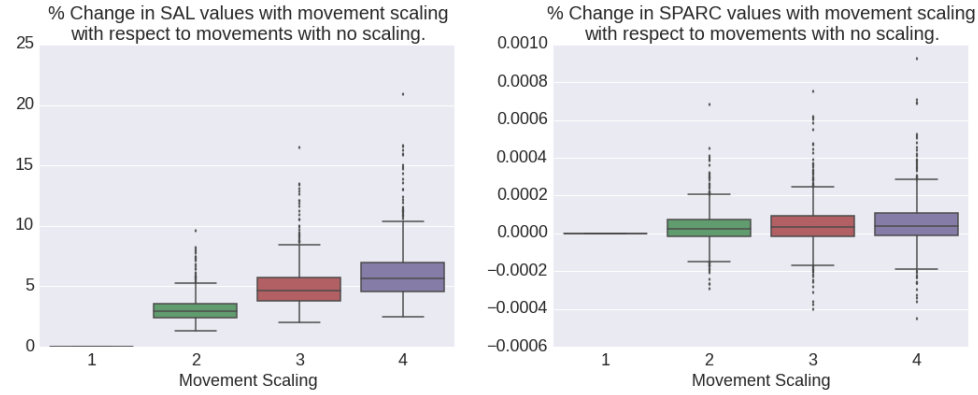


Figure 2.6: The effect of temporal movement scaling on SAL vs. SPARC [28]. SPARC is relatively impervious to movement scaling.

this, however, the amount of segments making up the total movement, N, should be known, and they should be at roughly equal lengths apart. These individual smoothness values can then be combined by use of a weighted mean function [28].

2.1.3.1.3 Minimum Jerk Research of velocity profile movements has been analyzed in the past for predicting hand trajectories and finding relationships to performance or skill [12, 29–31]. [29] suggests that movements in which more skill is involved result in a large increase in spatial variability of the trajectory, as well as a slightly increased peak velocity.

It has been observed in [3] that if coordination is modeled mathematically, the qualitative and quantitative features of the movement may be elucidated. This can be done through the defining of a cost function, which here is the square of the movement's jerk magnitude:

$$C = \frac{1}{2}\int_0^{t_f} \left(\left(\frac{d^3x}{dt^3}\right)^2 + \left(\frac{d^3y}{dt^3}\right)^2\right)dt. \tag{2.5}$$

When Equation 2.5 is minimized, the smoothest possible movement of the hand is theoretically obtained. This optimally smooth movement also depicts a movement in which only one submovement exists. This study showed that unconstrained point-to-point movements result in a relatively straight path, and a bell-shaped tangential velocity profile, as shown in Figure 2.7.

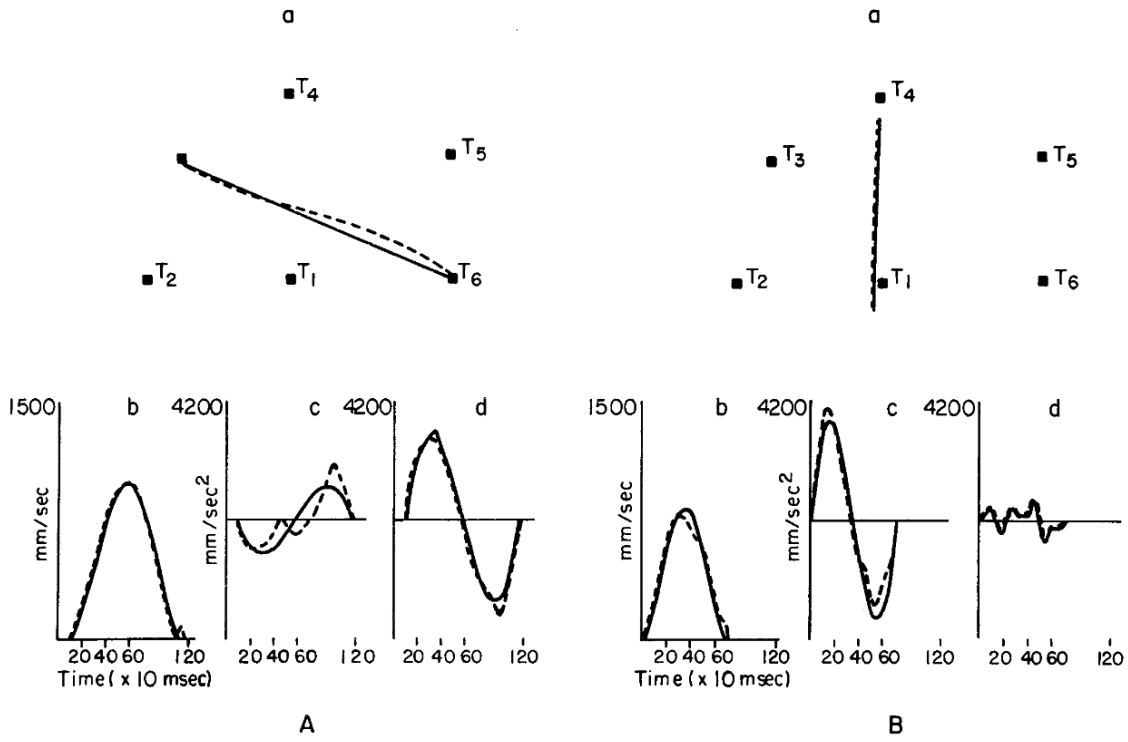


Figure 2.7: Predicted(solid lines) and measured(dashed lines) hand trajectories overlaid [3]. The minimum jerk model provides accurate predictions for the movements of recovering stroke patients.

Several models to predict and plot hand trajectories of unconstrained point-to-point movements exist, including the minimum jerk model, the support-bounded lognormal model, and the minimum snap model [25].

2.1.3.2 Objective

Given previous successes with neurophysiologically-derived models of hand movements, this thesis plans to test this with surgical data. This research will test the that the minimum jerk velocity prediction model will be able to accurately model the movements of expert surgeons, when compared to novice surgeons, using linear mixed effects modeling to account for variability of surgeons and task types.

2.1.4 Computational Skill Evaluation

Surgical skill evaluation is still a relatively niche field within research labs. Several different approaches exist with differing degrees of accuracy to compute surgical skill. The main approaches are using summary metrics calculated from surgical events and motion data, as well as statistical and machine learning models used from either kinematic data streams or video.

2.1.4.1 Neural Networks

Neural networks first were coined in 1943, when a McCullough theorized how neurons might work [32]. Since then, we have learned neurons do not work quite in this way but the term has stuck, due to the similarity to how the algorithms are computed. The first actual neural network was created in 1959 to eliminate echoes in phone lines [33]. Several groups have used neural networks(NNs) to help in the classification of images, videos, and other types of multidimensional data. There are several variations of neural networks, including:

- Convolutional NN (CNN)

- Recurrent NN (RNN)
- Gated Recurrent Unit (GRU)
- Long Short Term Memory Network (LSTM)

Neural networks use groups of 'neurons' or 'perceptrons', which make up one layer of a network, or a hidden layer, and feed data to remaining layers based on their activation function outputs. The input of a network is given to the neuron which computes one of several possible activation functions to assess the variablity of the inputs. Neural networks can be made of several layers, which are then called 'deep neural networks'. A neuron in the hidden layer will receive several inputs from different input neurons, and the output of that neuron's activation function is passed to the next layer. This continues until an output layer is reached, which will receive all the hidden layers' information and give an output based on that output layer's activation function. Figure 2.8 illustrates this basic architecture.

Different types of NNs can be used for different types of problems. CNNs are excellent at recognizing patterns that repeat across data. This is why they are mostly used for face and object detection. Instead of every neuron being connected to the following layer, however, in a CNN the first 'convolutional' layer takes a small portion of, for example, an image, as input, one by one. The filter runs through the entire image, and the results are restructured into an array. Once all the images have been fed through the network's pipeline, the algorithm is able to identify which objects are similar, and separates them into classes.

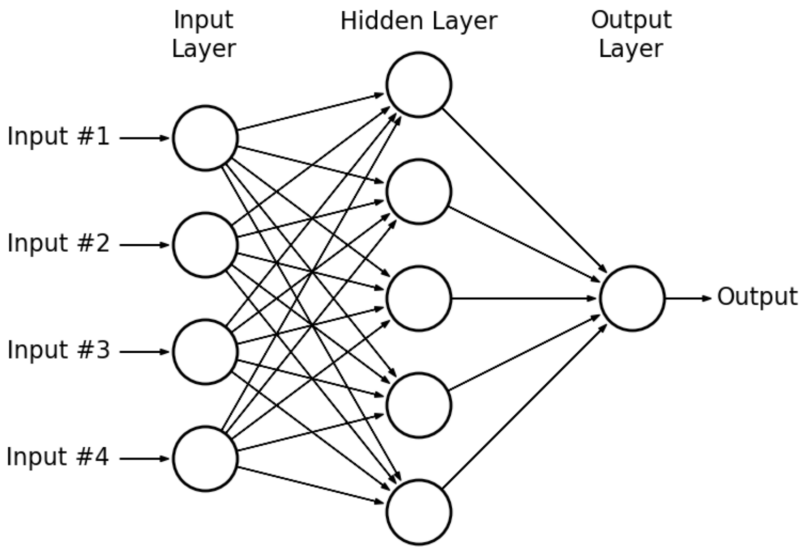


Figure 2.8: An example of the architecture of a neural network [34]. Inputs are sent through layers in which activation functions and array manipulations are used to arrive at a single output signal.

2.1.4.2 Recurrent Neural Networks

Recurrent Neural Networks (RNN) are different from CNNs in that they can take in arbitrarily-sized data and are useful for time-series information. This is possible because the output of the hidden layer in an RNN is fed back through the current layer with the previous layer's input. This allows RNNs to learn information from the past, as well as information from the present. RNNs are often used for language and vocabulary recognition, because they will 'remember' the context of a sentence, allowing prediction for certain words. RNNs are known to succumb to an issue known as the vanishing gradient problem. The vanishing gradient is a phenomenon susceptible to certain gradient-based optimization strategies in which a particularly small gradient is calculated during the optimization, and a domino-effect causes the algorithm to stop improvement due to no changes occurring in the gradient. Vanishing gradients can prevent a neural network from learning any useful information. This problem has been fixed, however, by long short term memory networks (LSTMs) [35].

2.1.4.3 Long Short-Term Memory

Long short-term memory networks are an even more recently created form of NN, which have longer dependencies than RNNs. In an LSTM, there are *two* inputs, a **cell state** and a **hidden state**. The cell state travels through a neuron to decide whether to dispose of the entering cell state or to store it. If the cell state is thrown away or forgotten, the hidden state makes an update to the cell state. The mechanism deciding whether to keep a cell state can vary between LSTMs, each having their pros and cons, but this is normally a mixture of tanh and sigmoid functions. An illustration of an LSTM cell is shown in Figure 2.9.

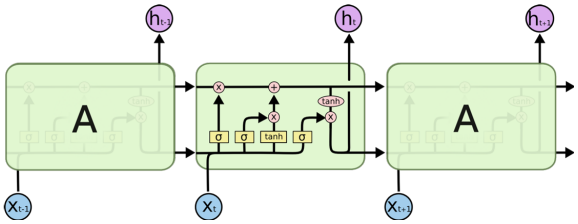


Figure 2.9: An example of the architecture of a long short-term memory network. A series of gates are used to create a sense of context for past mistakes and successes.

The most common way of setting up an LSTM is with a memory cell, input gate, output gate, and a forget gate. The terms are rather straightforward. The memory cell takes the input and stores it away, the input gate controls the likelihood of a new input flowing into the cell, the forget gate decides whether or not a value remains in the cell, and the output gate, controls whether the value in the cell is computed with the activation function. This activation function is usually the logistic function.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.6}$$

LSTMs have been used in speech recognition research with much success, and are part of the engine behind Apple's Siri speech recognition technology [36] [37], [38]. Speech models potentially provide promise for surgical skill evaluation due to the requirement of memory of past events. A speech model is required to keep a record of context to improve future predictions, just as a surgical skill classifier must record previous mistakes or successes to predict future skill levels.

2.1.4.4 Automated Performance Metrics

Automated performance metrics are summary statistics computed, most popularly with robotic surgery, as the daVinci surgical robot is able to track motion data throughout a surgical procedure. The daVinci robot, created by Intuitive Surgical (Sunnyvale, CA), also records events during a procedure, such as the number of times a surgeon moves their head in place of the console to conduct a movement, and the amount of times the hand or foot pedal is pressed on the console [39].

In [40] a combination of APM computation and task-evoked pupillary response (TEPR) was measured, to learn the affect of cognitive work load on the different skill levels. It was found that novice surgeons tend to move faster under high cognitive workloads, while experts move more slowly under the same cognitive experiences. They were additionally able to correctly classify the different skill levels using random forest classification with APM computation. It was found that the most reliable metric for calculating the skill of a surgeon was how often they used the camera clutching mechanism to move the camera while performing [41]. Other performance metrics exist relating to surgical skill which are discussed in more depth in Chapter 4.

2.1.4.5 Video Motion Data Analysis

Recent researchers have been able to use raw video data of tasks performed, and classify the skill of the performer [42–44]. These studies have used motion feature extractors from image data to discriminate different types of motion representative of certain skill levels. This research has been fairly accurate, but have only been tested with small datasets, and for subjects wearing brightly colored gloves so the video can

differentiate certain movements from others. In a real-world setting the performer would not be wearing these brightly colored gloves, making it much more difficult for image detection to identify hands and classify motions.

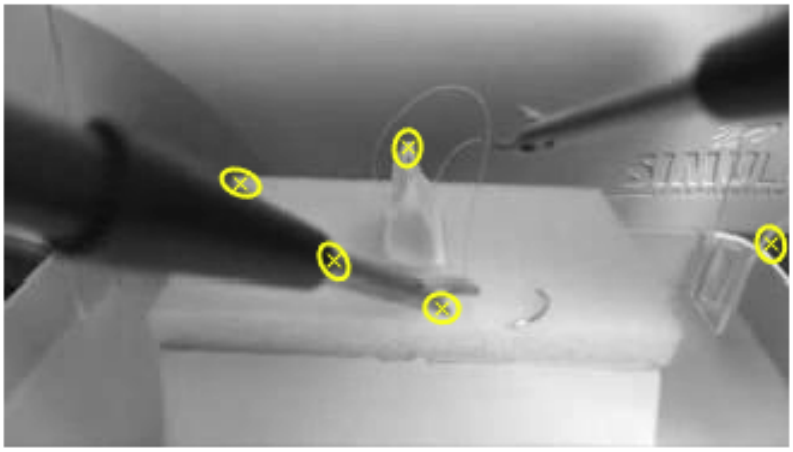
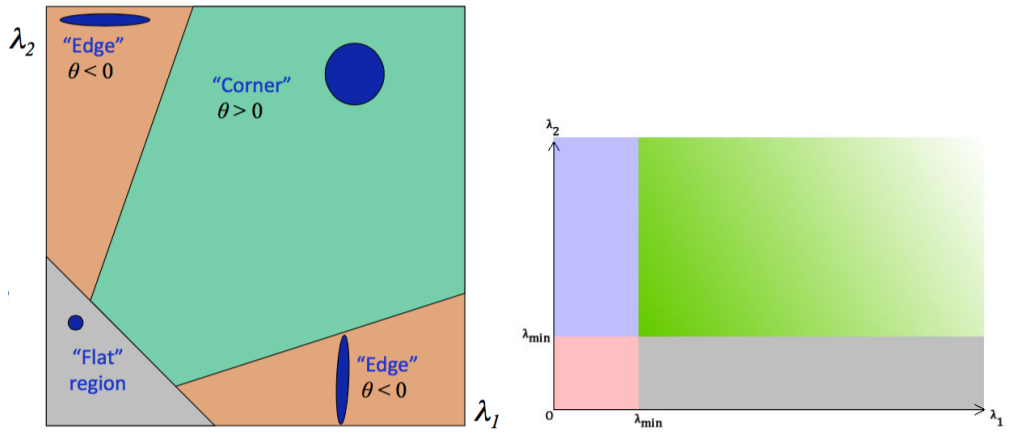


Figure 2.10: The detected SIFT points from a dry lab laparoscopic surgical task image (computed using MATLAB computer vision toolbox).

Several kinds of features may be extracted from video images. Some of the most popular methods are Harris corner detectors, Sped Up Robust Features (SURF), Scale Invariant Feature Transforms (SIFT) (shown in Figure 2.10, and, more recently, Spatio-Temporal Interest Points (STIP) [45]. These can all be used to help classify images in video, and are essentially small variations of basic corner detection [46], [47].

A more robust version of corner detection was created by Shi-Tomasi. This is a slight variation from Harris corner detection. Both methods calculate a score R, calculated for Harris detection by the equation



(a) The boundaries for harris corner detection to classify a point as a corner [49]

(b) The boundaries for Shi-Tomasi corner detection to classify as a corner [48].

Figure 2.11: Comparison of corner detection boundaries for Harris and Shi-Tomasi detection algorithms. Shi-Tomasi has a more rigid set of constraints, and will detect more corners because of these boundaries.

$$\begin{aligned} R &= detM - k(traceM)^2 \\ detM &= \lambda_1\lambda_2 \\ traceM &= \lambda_1 + \lambda_2. \end{aligned} \tag{2.7}$$

In Shi-Tomasi detection, the score is simply

$$R = min(\lambda_1, \lambda_2). \tag{2.8}$$

This small change has been shown to experimentally improve results. The change in boundary between classifying as a corner or an edge can be seen in Fig 2.11 [48].

These methods of extracting features consist of detecting corner points and subsequently computing histograms of oriented gradients (HoG) in three directions (two

spatial and one temporal direction) [50]. These 3D HoG features will provide information for how the objects move through time as well as space. Several researchers have used variants of this technique for surgical skill evaluation [51–54]. One study went a step further and used these features to create frame kernel matrices.

A standard kernel matrix is an equation which gives a measure of similarity between two matrices. Using this with the features in image frames conveys the similarity between image classes in subsequent frames. An example frame kernel matrix is shown in Figure 2.12. This approach used image features to divide the motions into 5 distinct motion classes using K-Means Clustering, and subsequently classify surgeons by which motion class most of their movements belonged [55]. This most publication reported an accuracy of 88%, making it the benchmark for other publications to compete with.

2.1.4.6 Lack of Quality Data

Most previous research in skill evaluation has suffered from a lack of robust data sets. The main studies completed with automated performance metrics were conducted with 26 cases. The most promising work which used video analysis of motion features used a data set of just 18 participants [56]. A summary table for some other popular studies with the size dataset are shown in Table 2.3. In order to create reliable models, studies with data of larger volumes must be used and made public to the research community so that comparable results may be communicated.

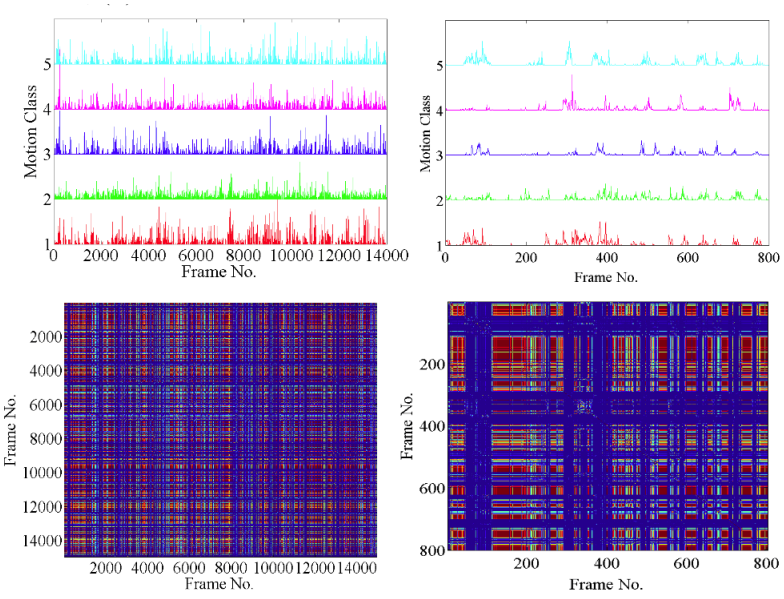


Figure 2.12: Top: Discriminated motion counts corresponding to each class of STIPs. Bottom: Frame kernel matrices example, displaying the relationship between various segments of a video [55].

Group	Technique	N
Chen(2018) [39]	APM	70
Nguyen(2019) [40]	APM	26
Zia(2017) [56]	Motion Analysis	18
Fard(2018) [57]	Motion Analysis	3
Sharma(2014)	Video Analysis	16
Zia(2017) [53]	Video Analysis	8
Funke(2018) [58]	Video Analysis	8
Wang(2019) [59]	Video Analysis	8

Table 2.3: Summary of popular skill evaluation studies, and lack of high volume data.

2.1.4.7 Objective

Given previous successes in fields unrelated to surgical skill evaluation in which time is a large part of the equation in correctly making predictions, it is believed an LSTM could be used in surgical settings. This thesis aims to create a machine learning model which uses an LSTM with tooltip kinematic data from dry lab laparoscopic training procedures to make predictions which will succeed in assessing the skill of novice and expert surgeons, surpassing previous research efforts summarized in this chapter.

2.1.5 Biological Motion Perception

As mentioned earlier in this chapter, task time seems to be the summative metric which most often correlates to surgical skill, as determined by experts or crowd scores of video evaluation [13]. As these are relatively subjective forms of measurement, they can be riddled with inaccuracies and biases that humans subconsciously obtain. It is possible that something which occurs in the brain causes humans to unconsciously reward quicker speeds when they are unaware of the speed up.

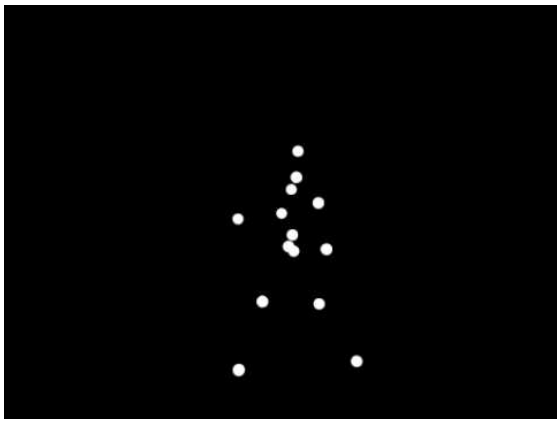


Figure 2.13: Example of a point light walker. Each of the dots represent a limb or joint of the human body.

2.1.5.1 Visual Stimuli Research

Psychophysical experiments have tackled the question of sensory stimulus using different methods of manipulation. The first of these is a study in which animations of point light (PL) animations were used, illustrations of moving dots representing human joints (Fig. 2.13) [60]. This concluded that observers can classify gender from one of these animations with 63% accuracy. Additionally, manipulations of unnatural arm swinging, different walking speeds, as well as removal of different joints significantly reduced the ability to recognize gender, to nearly 50%. Interestingly, faster walkers did appear more female to observers, although to a non significant degree [61]. This work was continued in a similar study of PL animations, finding that two revolutions of gait are needed to accurately determine gender [62], and another finding that only one-tenth of a second is needed to recognize human motion [63]. This work shows that motion perception is somehow linked to dynamics. Our brains appear to have an ability to effortlessly decode this information.

2.1.5.2 Response to Abnormal Dynamics

It has been shown when these various animations are manipulated to move at abnormal or unnaturally slow speeds, instead of perceiving slower movement, observers perceived the animation to be rotating [64]. In contrast, when animations are played at speeds outside the realm of human possibility using spatiotemporal filters, observers also become inefficient at recognizing the gender [2]. Likewise, [65] found that humans are able to accurately assess which speeds appear natural. It is possible that biological motion perception has a speed dependence in which the speed occurring must appear to be realistic for the motion perception unit to properly process it.

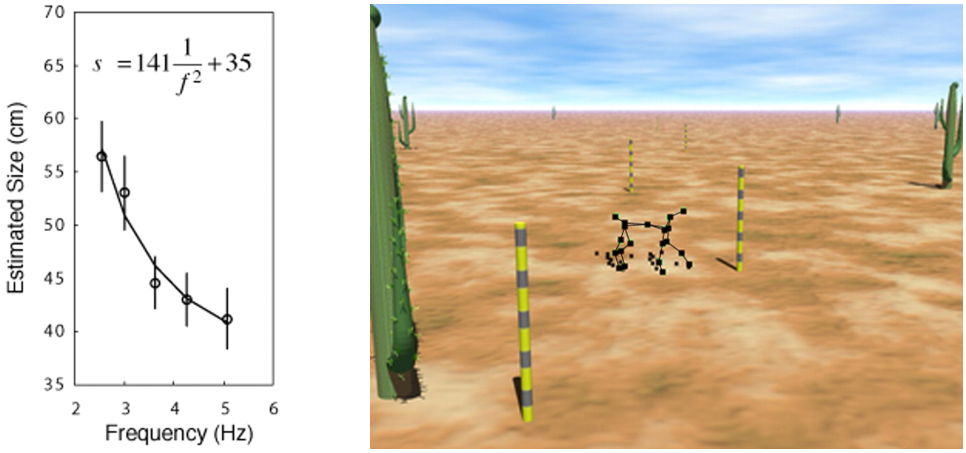


Figure 2.14: Illustration of a PLW representing an animal (right) and the change in perceived size as the gait increases (left). This suggests a link between speed and biological motion perception [66].

2.1.5.3 Objective

Driven from previous research in biological motion perception, it is hypothesized that humans will perceive the skill of a surgeon differently when video playback speed is increased. There may exist a biological motion processing unit in the brain which is able to detect movement which seems 'unrealistic' when moving at speeds too fast to

attribute to normal human speeds. Once this threshold is met, I hypothesize the brain will begin to control for the increase in speed and this perception of increased skill will cease. This infers a bell-shaped unimodal function of perceived skill as video playback speed is increased.

2.1.6 Datasets to be used in this work

2.1.6.1 Basic Laparoscopic Urologic Study

The Basic Laparoscopic Urologic Skills(BLUS) dataset is made up of 535 videos recording four different laparoscopic tasks: peg transfer, pattern cutting, suturing, and clip applying. This data was gathered from 8 urologic training centers in the United States from 76 unique subjects who ranged from 1st to 5th year residents. There are 49 R1 subjects, 11 R2 subjects, 6 R3 subjects, 6 R4 subjects, and 4 R5 subjects. The videos have: 133 peg transfer, 127 pattern cutting, 135 suturing, and 140 clip applying tasks. These data include both videos of the task and tool motion data. Included are also expert and crowd-sourced ratings for each trial [19]. Frames from three of the four tasks are shown in Figure 2.15, with the range of scores for each of the four tasks in the dataset shown in Figure. 2.16.

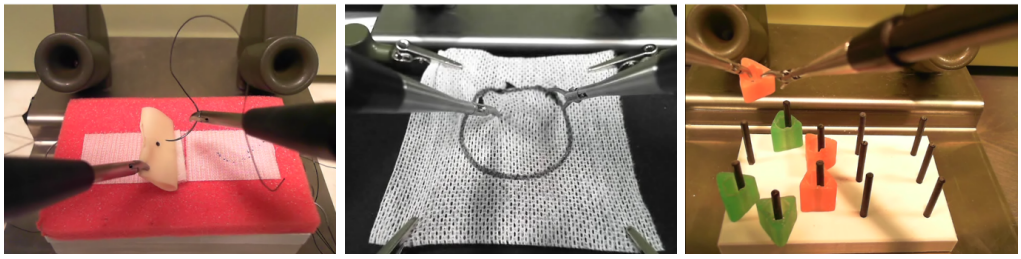


Figure 2.15: Frames from a (left to right) suturing, cutting, and peg transfer task in the BLUS dataset.

2.1.6.2 Robotic Surgery Readiness Study

The Robotic Surgery Readiness Study data set is made of 343 videos of live robotic surgery cases which all hold accompanying kinematic tooltip data logging the tool motion during each video. This data set arose through research aiming to test whether

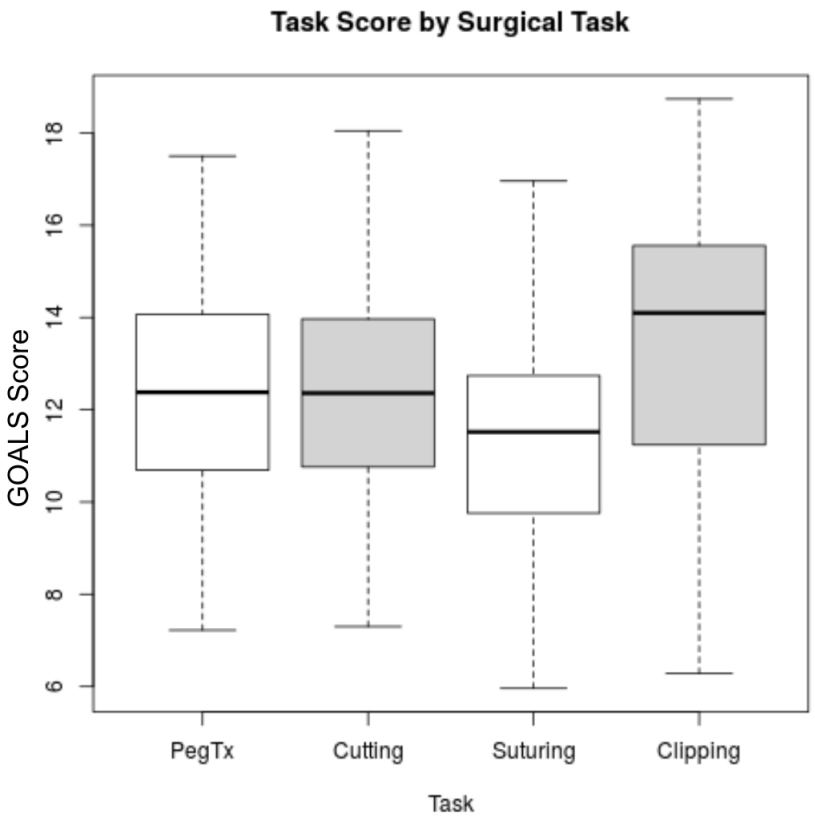


Figure 2.16: Range of crowd-assessed surgical skill evaluation mean scores for each of the BLUS tasks. On average clipping tasks receive higher scores than, for example, suturing tasks.

pre-operative warm-up using a validated simulation curriculum improved robotic surgical performance among practicing surgeons with lapses in practice. These data are explained in much more detail in Chapter 6 as they were created during the past three years. An example figure from the RSR dataset is shown in Figure 2.17.

2.1.7 Linear Mixed Effects Modeling

When data is grouped into several subgroups in which a relationship may exist, a useful mechanism to use is a linear mixed effects (LME) model. The assumptions made

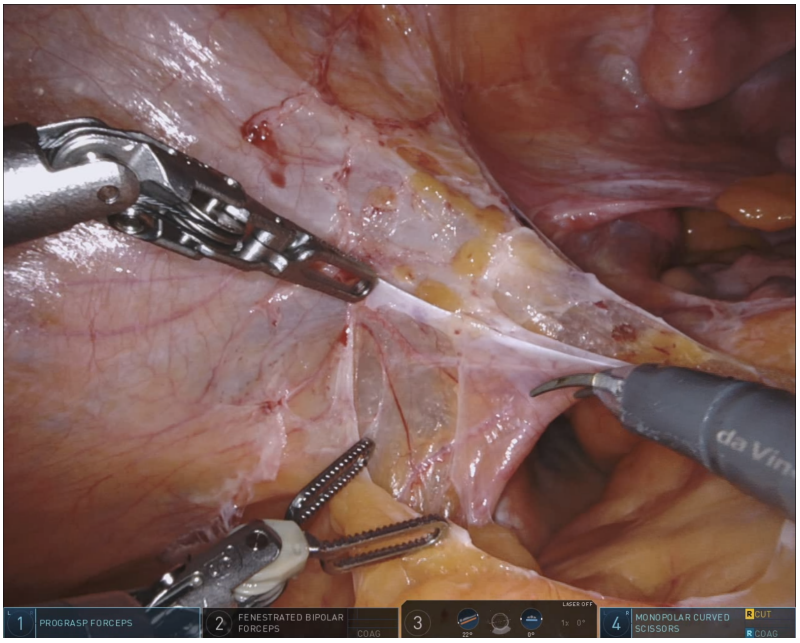


Figure 2.17: Frame from a video in the RSR data set, recorded using a daVinci surgical robot (Intuitive Surgical (Sunnyvale, CA). Typical robotic surgeries recorded were prostatectomy, hysterectomy, and partial nephrectomy, with all procedures being recorded in Washington state.

before using an LME model is that the data is not independent and thus we can use a combination of fixed and random effects to create an advanced linear regression which will explain the hierarchical structure of the data. One example of this sort of data may be several doctors seeing a variety of patients who either have or have not taken a drug [67].

Fixed effects are parameters that do not vary throughout the dataset, and are assumed to be measured without error. This could be something like the amount of a drug taken for specific groups of patients. These variables are easily repeatable in multiple studies. In contrast, a random effect is something that changes throughout the experiment, such as a separate doctor seeing each patient. In this way we can account

for different doctors having different baseline levels at which they may make certain decisions, as well as each doctor having a different range of symptoms they would like to see before making certain decisions. Modeling random effects in this example gives each doctor a specific slope in the regression, and each fixed effect is given a specific intercept in the regression. By modeling our data this way we can account for the variability of fixed effects, things measured without error, but also have variables which account for the factors which changed throughout the experiment. This general setup in mathematical form would be

$$y = X\beta + Zu + \epsilon \tag{2.9}$$

in which the outcome y is a vector of size Nx1, X is a Nxp matrix of the p predictor variables, β is a px1 column vector of the fixed-effects coefficients, Z is the Nxq design matrix for the q random effects, u is a qx1 vector of the random effects, and ϵ is a Nx1 column vector of residuals, or errors not explainable by the model [68].

When testing a linear mixed effects model, in order to learn whether a certain parameter was significant in the regression, the standard method to use is an ANalysis Of VAriance (ANOVA) test. The typical setup is to use the null model, the model which does not include the parameter of interest, and test with a model which includes an effect for the desired parameter. The information criteria from an ANOVA test can then give a sense for whether the parameter is significant in predicting an outcome by comparing the information criteria. If the information criteria for the new model is lower than the original, as well as having a reasonable p-value, this confirms the parameter is significant in predicting the outcome variable.

Chapter 3

Neurophysiological Models

This chapter provides the methods used to test whether velocity prediction models are correlated to surgical motion segments.

3.1 Testing the Relationship of Minimum Jerk Model to Surgical Skill

3.1.1 Methods

The already obtained velocity trajectories from the BLUS dataset were segmented using an algorithm which tracked when the tooltip velocities fell to near zero. Additional forms of segmentation have been used, including presence of a tool grasp in each motion, and motions corresponding to the force of a grasp passing a certain threshold, but the most statistically significant form of segmentation is for segments separated at near-zero velocity segments.

The speed segments from each of the performances in each of the four BLUS tasks

were fit to a minimum jerk velocity profile, and the mean squared error of the difference between the points on the fitted curve and the raw data were recorded for each movement and recorded as belonging to the surgeon who made that specific movement. In an effort to make the data more manageable, these error values were used for the remaining calculations.

Initially a fixed effects model was used, which accounted for a fixed effect in each β coefficient, but did not account for the error of each specific surgeon. The fixed effects model was of the form

$$TaskScores \sim \beta_1 + \beta_2 + \beta_3 + \beta_4 + MSE * Subjects, \tag{3.1}$$

in which each of the β values are coefficients used to fit the minimum jerk model to the specific hand movement, and MSE is the mean squared error of the fitted line to the model. Additionally a linear mixed effects model was used which included a random effect for each of the subjects and tasks. An ANOVA test was then used to compare the null model to a model which included the error term. The model being tested was

$$TaskScores \sim (1|Subjects) + (1|Task) + MSE. \tag{3.2}$$

This equation allows for variability within Tasks as well as within specific performers of the tasks. This allows the model to know each task will have a different range of scores, as well as knowing each surgeon will have a baseline level of performance.

3.1.2 Results

Using a surface plot which allows for plotting with various values of the alpha transparency coefficient, the velocity of each segment of each performance of Peg Transfer

tasks was overlaid onto one figure, with the time component normalized. As can be seen in Figure 3.1, a vague shape emerges from the overlaid segments showing the general trend of grasping movements. Figure 3.2 shows a visualization of the segmented grasps for the first observation of the peg transfer tasks, with normalized time and velocity segments.

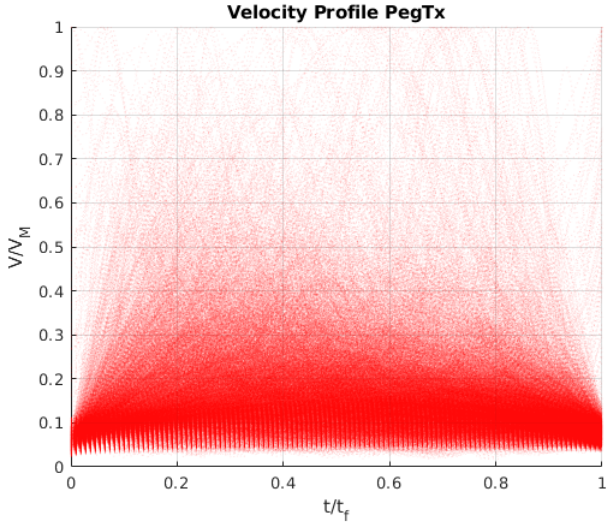


Figure 3.1: All Peg Transfer task segments overlaid to show the general shape of peg transfer task grasp segments. One single line represents one single movement by a surgeon. Movements were defined as starting and ending when the hand reached nonzero speed. Time is normalized by the duration of each segment and velocity is scaled by the peak velocity magnitude of that segment.

Figure 3.3 shows an example of the difference between the minimum jerk model's fit to grasps and the raw grasp data. Figure 3.4 show the results of the fixed effects model in Equation 3.1, with the actual vs. predicted scores plotted. These show a correlation of 0.521.

This result for the model in Equation 3.2 is shown in 3.5 and has a correlation of 0.826 to the outcome scores. To test the viability of this model, an ANOVA test was used

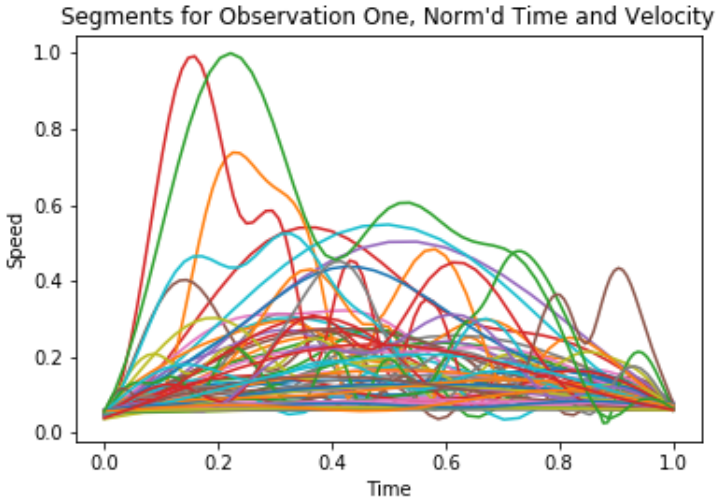


Figure 3.2: Example of segmented grasps, from the first performance of the peg transfer tasks only. Showing the data in this manner allows seeing the variability that exists throughout the different movements in a surgical task performance.

to compare the linear mixed effects model shown above, to the null model, excluding MSE, as that is the variable of interest. The results are shown in Table 3.1. The AIC, or Akaike Information Criterion, is a well-known measure of the relative goodness of fit of a model, when being compared to a similar model. This AIC can be calculated by

$$AIC = 2k - 2ln(\hat{L}) \tag{3.3}$$

in which k is the number of parameters, and $\hat{L}$ is the maximum likelihood. When comparing two or more models, the model which has a lower AIC is preferred. It is standard practice to use AIC to compare a model to the null model, which excludes the variable of interest.

As can be seen, the AIC is lower for the model with an MSE term included, and the χ^2 value is 951.45 with a p-value of ≤ 0.0001. From the summary of the LME model,

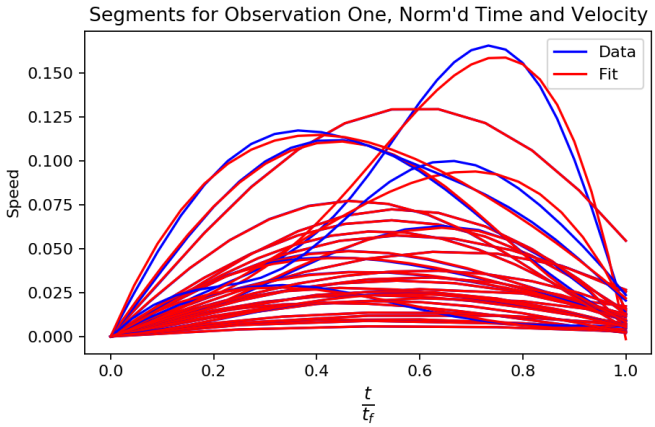


Figure 3.3: Unimodal movements from peg transfer tasks fitted to minimum jerk movements described in Chapter 2.1.3.1.3.

the MSE fixed effect term raises the predicted score of a surgeon by 0.4324 ± 0.2803. This suggests a higher error actually leads to a higher score, or that higher scores mean a worse fit to the minimum jerk model.

The LME model, plotted separately into the four tasks performed, can be seen in Figure 3.6.

Model	AIC	BIC	logLik	deviance	Chisq	Chi Df	P-Value
Null	74390.18	74421.87	-37191.09	74382.18	NA	NA	NA
MSE Model	73448.73	73520.03	-36715.36	73430.73	951.4489	5	≤ 0.0001

Table 3.1: ANOVA test results for the mixed effects model on the minimum jerk viability as a predictor variable for task score value.

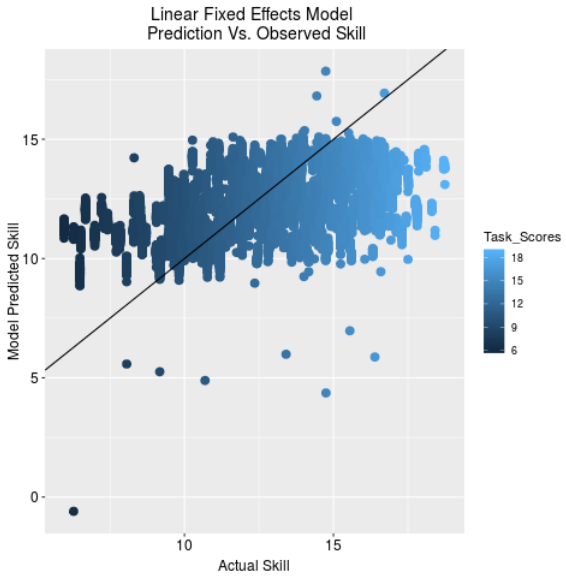


Figure 3.4: Fixed effects model of observed vs. predicted skill, for all BLUS tasks combined, using minimum jerk model, subject and task identity, in which the fixed effects accounts for a constant intercept for all of these factors. The fixed effects model behaves quite poorly.

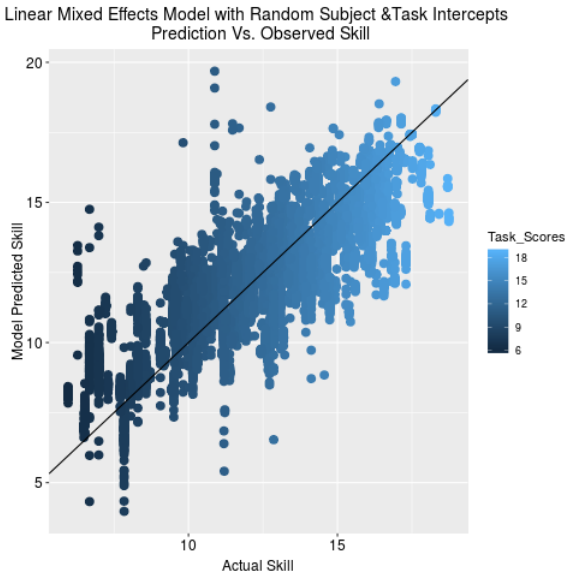


Figure 3.5: Linear Mixed effects model's predicted vs. observed skill using minimum jerk model fit error, task, and subject indentity, in which this model accounts for varying intercept values for each task, and subject. This model is able to predict outcomes much better than the fixed effects model, with a correlation between actual and predicted skill of 0.826.

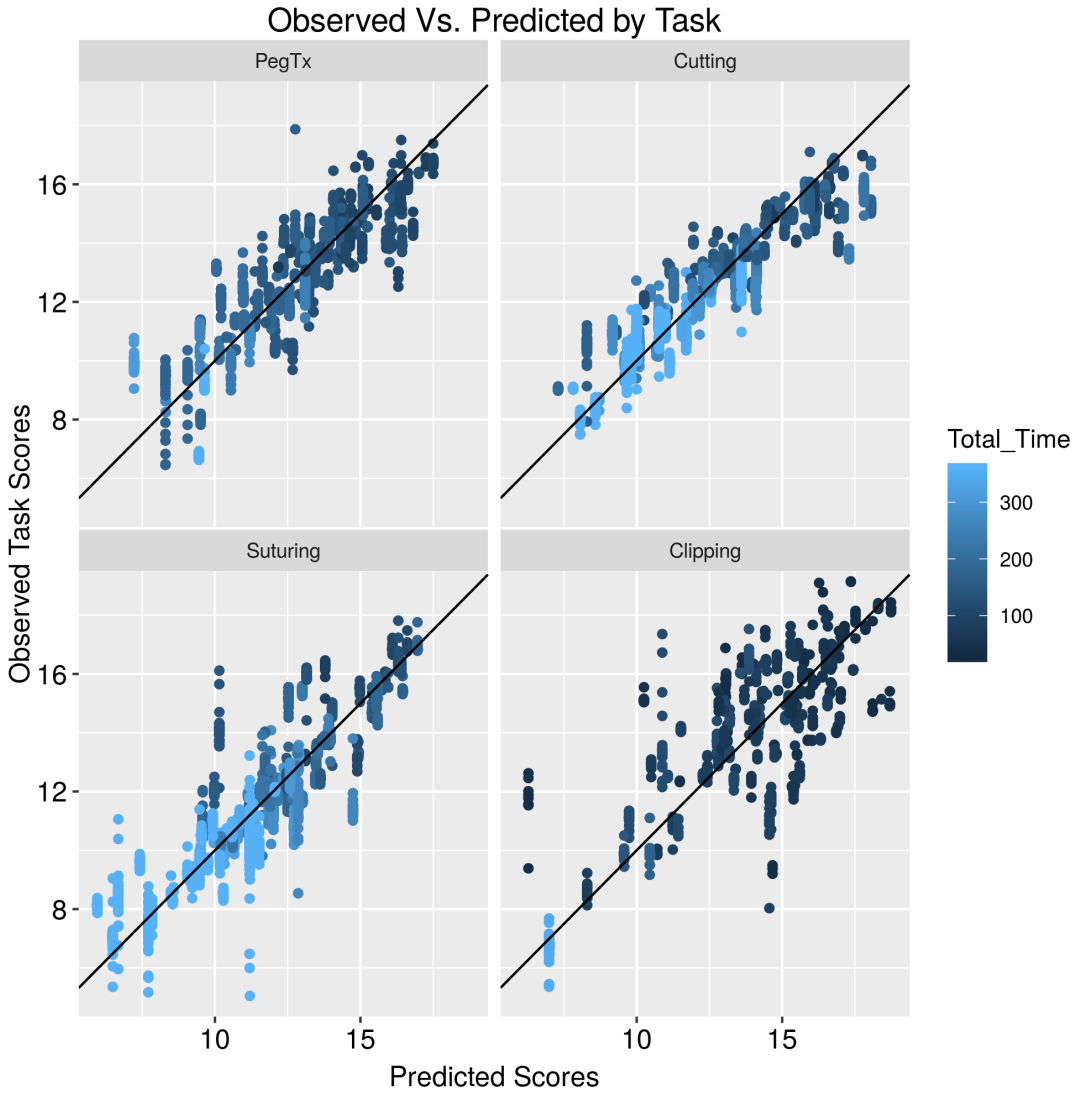


Figure 3.6: LME predictions grouped by task, with colorbar illustrating total time for task completion. There is a strong correlation between observed task scores and predicted scores, with slower performances usually receiving poorer of both.

Chapter 4

Computational Skill Evaluation

This chapter will discuss efforts in creating a machine learning model for use with multidimensional kinematic data extracted from the BLUS dataset. This model was used to classify the skill of surgeons performing various laparoscopic training procedures. This is a reproduction of a journal paper submitted to the International Journal of Computer Assisted Radiology and Surgery in March 2020 [69].

4.1 Bidirectional Long Short-Term Memory for Surgical Skill Classification of Temporally Segmented Tasks [69]

4.1.1 Abstract

Purpose: The majority of historical surgical skill research typically analyzes holistic summary task-level metrics to create a skill classification for a performance. Recent advances in machine learning allow time series classification at the sub-task level, allowing predictions on segments of tasks, which could improve task level technical skill assessment.

Methods: A bidirectional long short-term memory (LSTM) network was used with 8 second windows of multidimensional time-series data from the Basic Laparoscopic Urologic Skills (BLUS) dataset. The network was trained on Experts and Novices from four common surgical tasks. Stratified cross validation with regularization was used to avoid overfitting. The misclassified cases were re-submitted for surgical technical skill assessment to crowds using Amazon Mechanical Turk to re-evaluate, to analyze the level of agreement with previous scores.

Results: Performance was best for the suturing task, with 96.88% accuracy at predicting whether a performance was an expert or novice, with 1 misclassification, when comparing to previously obtained crowd evaluations. When compared with expert surgeon ratings, the LSTM predictions resulted in a Spearman coefficient of 0.89 for suturing tasks. When crowds re-evaluated misclassified performances it was found that for all 5 misclassified cases from peg transfer and suturing tasks, the crowds agreed more with our LSTM model than with the previously obtained crowd scores.

Conclusion: This technique shows promise both for classifying performances of surgical tasks and evaluating sub-task level segments of performances. This method could be used for near real-time feedback to surgeons. Further work will be done to

expand this methodology to a larger cohort of surgical performances including robot-assisted surgery.

4.1.2 Introduction

Computationally assessing the skill of a surgeon in an objective manner using tool motion has proven a complex problem with many challenges. Previous research has relied mostly on summary performance metrics from kinematic data [41], [13, 27]. Unfortunately, these metrics typically failed to completely discriminate novices from experts, that is to never misclassify "obvious" novices vs. "obvious" experts – the so-called Minimally Acceptable Classifier (MAC) Criterion [70]. Recent advances in machine learning techniques have expanded the possibility of using time series datasets to evaluate surgeries and surgical tasks [71].

The de facto gold standard for determining the level of skill in a surgical performance is video-based evaluation by an expert surgeon [72]. Skills assessment is normally evaluated with the use of one of several possible established assessment schemes that utilize Likert-scale scoring of subdomains relevant to skill, such as bimanual dexterity, or tissue handling. An example of one of these assessment methods, the Global Operative Assessment of Laparoscopic Skills (GOALS) evaluation scheme is shown in Table 2.1 [24]. This is both a time-consuming and laborious process for experienced surgeons whose valuable time could rather be spent performing life-saving surgeries. In recent years it was found that crowd based evaluation of surgical performance is able to achieve the same standard of accuracy in predicting technical skill in surgery, with a 100% pass/fail rate, as compared to expert surgeons [22]. This may still be a subjective measure of technical skill and prone to bias [73], thus an objective technique to computationally evaluate the technical skill of a surgeon would be beneficial.

Long short-term memory networks (LSTMs) are an adaptation of recurrent neural networks which are capable of analyzing past events in a time series to learn how they might affect a present time index [35]. This is possible through a series of gates in the architecture of the network which hypothesize, learn, and forget predictions by deducing what information to ignore and which to emphasize in the training process. These networks have been further improved through the implementation of bidirectional networks [74]. These behave by training a network in the forward direction, while also training a network in the reverse direction, thereby connecting two hidden layers in different directions to create one output from twice as much information, and allowing the LSTM to use this increase in information to achieve better results.

The objective of this investigation was to evaluate the importance of temporal segmentation of surgical tasks for quantifying the skill of a surgeon. The hypothesis of this work is that experts don't behave in an expert-like manner throughout the entirety of a task, and likewise for novices, but instead that the main factor in deciding upon a surgeon's overall technical skill is the number of expert-like to novice-like segments in footage of a surgical task, and that bidirectional LSTMs have the ability to learn this information from kinematic tool motion data.

4.1.3 Methods

4.1.3.1 Dataset

This study used the Basic Laparascopic Urologic Study (BLUS) dataset, described in detail in [19], with a summary re-iterated here for convenience. This dataset arose

from a gap in the field, in which no educational surgical certification process existed for urologic surgery, as opposed to how the Fundamentals of Laparosocopic Surgery (FLS) exists for general surgical procedures [75–77]. The BLUS training curriculum aimed to address urology appropriate skills improvement by recording video performances in an initial validation project of over 450 videos [78].

This dataset contains 454 videos of surgical performances consisting of four surgical tasks (110 peg transfer, 110 pattern cutting, 115 suturing, 119 clip applying), which are performed by medical students, urology residents, fellows and faculty surgeons from eight academic urology training centers in the United States [79]. Each trial of a surgeon performing one of the four tasks was recorded at 30 fps with a fixed camera-position of the laparoscopic tools interacting with the training field. Each trial additionally has kinematic data, sampled at 30 Hz, logging the tooltip positions, grasping force, and the jaw angles during the performance, as well as demographic information for each performer being obtained. A GOALS score was obtained for each video via crowd evaluation, and an expert evaluation was obtained for a subset of videos.

Previous research regarded suturing the most clinically relevant of the four tasks, as these performances require the mastering of needle and suture handling which are more similar to what is encompassed in real surgery vs. transferring synthetic blocks or gauze cutting. However, all four tasks were used in this study in an attempt to provide a classification scheme which can successfully separate experts from novices. Here a novice was defined solely as an "obvious novice", or someone who should never be allowed to operate and experts solely as "obvious experts", or surgeons who should never be disqualified from operating [70]. An obvious expert was chosen such that the performer was in the top 20% of previously obtained GOALS scores for that particular

BLUS task. The obvious novices were chosen in the same fashion such that they were in the bottom quintile of these domains. This method aimed to provide two well discriminated clusters of skill levels that should demand no misclassifications.

4.1.3.2 Skill Classification from Temporal Segmentation

4.1.3.3 Data Partitioning

In the preliminary analysis, a sliding window data partitioning technique was used in which windows of varying sizes with varying overlapping lengths were tested to find the optimal parameters. The network was trained by using the novices and experts from the top and bottom quintile in each of the four tasks such that the classifier could more easily find separating features to define each class. These values are shown in Table 4.2, where the minimum possible GOALS score is 4, and the maximum is 20. These labels are then converted into binary variables based on whether they are considered an expert or novice. This results in 16 expert videos and 16 novices videos for peg transfer, suturing and cutting tasks, and 16 expert videos and 17 novice videos for the clip applying task.

4.1.3.4 Window Parameter Selection

Due to the relatively small dataset resulting from sub-selecting only obvious novices vs. obvious experts, a stratified cross validation scheme was computed on the experts and novices, in which two performances were left out of training. Each of these groups of two performances consisted of an expert and a novice, in an effort to keep the training response values at equal proportions and prevent overfitting of the dataset. In addition, to avoid oversampling of novice performances, as novices consisted of about 70% of the

total data, being that novices usually take longer to perform, the expert-labeled performances were sampled with a step size half as large as the novice step size, to simulate more training data. This sliding window and the specified sampling control technique is illustrated in Fig. 4.1.

4.1.3.5 LSTM Parameters and Architecture

The bidirectional LSTM network consisted of a one layer 32 unit bidirectional cell, followed by a 50% dropout layer, and a sigmoid activation function. L2 regularization was also used to further prevent overfitting. Each network was trained to 100 epochs using a binary crossentropy loss function, and the Adam optimizer. Each segment of a test performance was fed to the network and evaluated. After all segments were evaluated, the LSTM outputs a probability of how likely a segment is to be expert-like vs. novice-like. The task as a whole was considered an expert level performance if the mean of the predictions resulted in a prediction of greater than 0.50, and vice versa.

After the initial results were obtained from using experts and novices as training and test data only, additional testing was done using intermediates as test data on these models. During this trial, the same training technique was used as discussed previously, this time alternatively using intermediate level performers as the test data. This training was repeated ten times per validation set, with different random number seeds being assigned at the intialization of each trial. For each intermediate performance prediction, the ten different predictions were averaged for each segment of the performance. An example of this is shown in Fig. 4.2.

Using these intermediate performance prediction values, it was now possible to get a

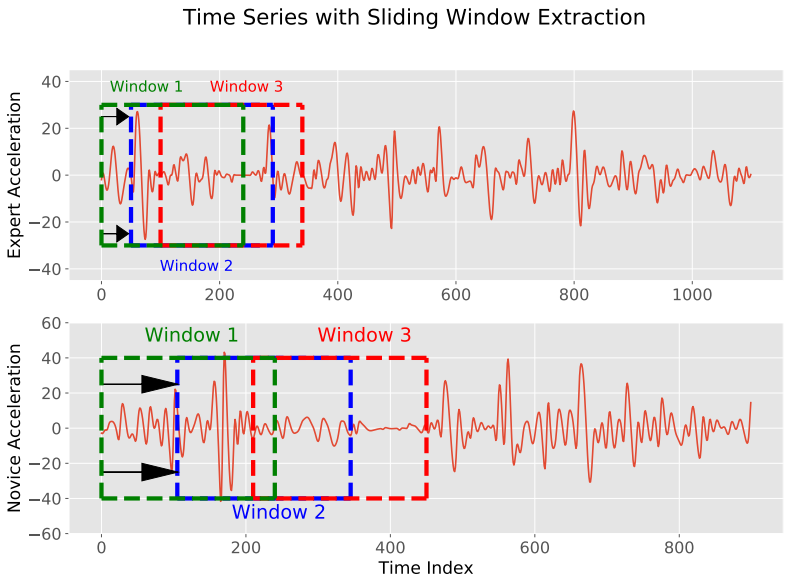
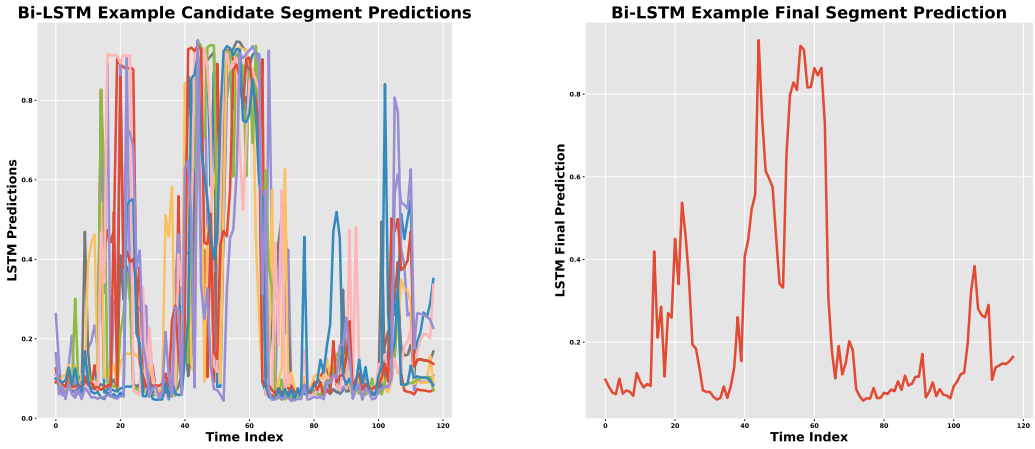


Figure 4.1: Example left hand tool acceleration time series with the sliding time window extraction method applied to both a novice series and an expert series. A novice performance is indexed with a window overlap of 105 time indices, and an expert performance is indexed with a window overlap of 50 time indices.

rough approximation of the correlation between the previously obtained GOALS scores and the average of the performance's LSTM predictions. These values were obtained for the experts and novices, as well as the intermediate performers.

4.1.3.6 Crowd Reassessment

As suturing and peg transfer tasks are agreed to be the two most separable and clinically relevant tasks from the BLUS dataset, these two were chosen to have their misclassified videos reassessed by crowd workers. In addition 5 obvious experts and 5 obvious novices from each task were randomly selected for reassessment. Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each



(a) 10 Test Predictions for a suturing performance.

(b) Final averaged prediction with outliers removed.

Figure 4.2: Process of computing final suturing performance prediction. All ten candidate predictions of a performance are averaged at each index to compute the final prediction.

non-expert crowd worker was paid an average of $0.50 to watch and evaluate the short video of the surgical task. The goal was to compensate the evaluators at a rate of approximately $10/hour. A user interface was created which asked crowds to rate the skill level of the performance using the GOALS assessment method, as obtained previously. These videos were given to crowds one at a time, using 40 crowd workers per video. The mean of the ratings for each video was then taken and compared to the previously obtained ratings for the misclassified videos, to find the level of agreement.

4.1.4 Results

4.1.4.1 Skill Classification from Temporal Segmentation

After performing a grid search with different hyperparameter values, the network was found to perform optimally with a window sample size of 240 time indices, approximately equal to 8 seconds, with a 3.5 second overlap at each window for novices and

Dropout	Batch Size	L2 Regularization
0.1	120	0.1
0.2	**240**	**0.2**
0.4	360	0.3

Table 4.1: Main hyperparameters tested during model evaluation in Section 3.1, with most optimal results in bold.

approximately a 1 second overlap for expert performances, which results in the number of samples from experts and novices to be roughly equal. Most of the values tested are shown in Table 4.1.

Table 4.2 illustrates the accuracy of these networks in classifying overall skill in surgical task settings, achieving over 96.88% accuracy for the suturing task, which is usually seen as the most clinically relevant task. Table 4.2 also reports the expert and novice specific accuracies, showing that all of the novices for suturing were correctly classified. Of the 32 suturing videos labeled as experts or novices, only one was mislabeled. By getting the average of all predicted segments in a specific performance, the algorithm results in a number between 0 and 1, codifying the LSTM's prediction of skill for the performance. These predictions can then be used to arrive at a correlation coefficient signifying the degree of correctness in the neural network at evaluating intermediate performers in addition to experts and novices, when comparing to crowds, illustrated for each BLUS task in Fig. 4.3. Correlation values for each task are in Table 4.3. Suturing had the highest Pearson correlation coefficient with a value of 0.86.

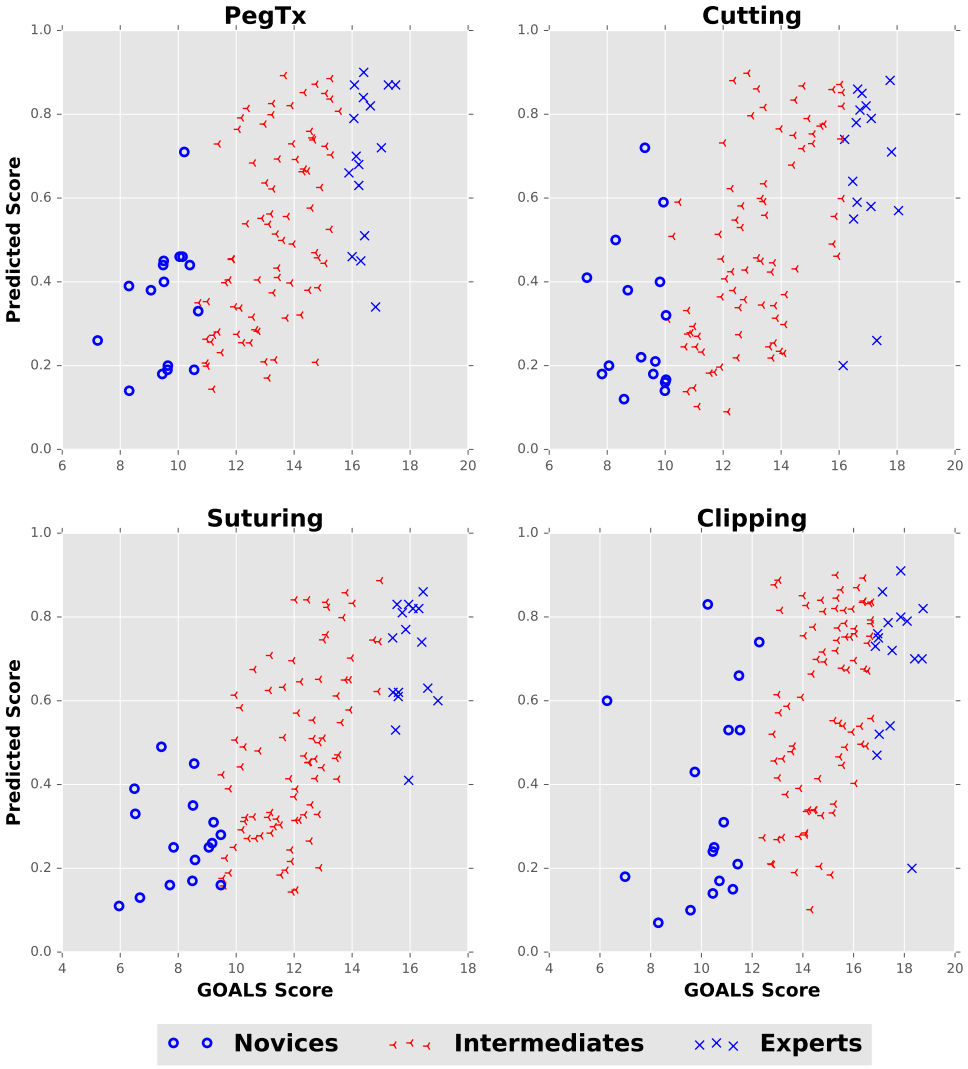


Figure 4.3: The LSTM's prediction for all performances from each of the tasks in the BLUS dataset. The LSTM was only trained on experts and novices (summary score labels with cross validation), from Section 3.1.

The suturing task was the most accurate of the four tasks in the BLUS dataset. Twelve of the suturing performances were previously rated by expert faculty surgeons in addition to being evaluated by crowds. Fig. 4.4 shows the faculty scores plotted against the skill predictions obtained from the bidirectional LSTM. As can be seen,

Task	Accuracy	Expert Threshold	Novice Threshold
Suturing	**96.88%**	15.40+	9.47-
Peg Transfer	87.50%	15.89+	10.69-
Cutting	87.50%	16.14+	10.03-
Clip Applying	73.33%	16.86+	12.28-

(a) Accuracy for each BLUS task's experts and novices, from Section 3.1.

Task	Novice-Specific Acc.	Expert-Specific Acc.
Suturing	**100%**	**93.75%**
Peg Transfer	**93.75%**	81.25%
Cutting	87.50%	87.50%
Clip Applying	68.75%	87.50%

(b) Novice and expert specific accuracy for each BLUS task, also known as sensitivity and specificity.

Table 4.2: Accuracy results computed by the bidirectional LSTM for each BLUS task.

Task	Spearman	Pearson
Suturing	0.76	**0.86**
Peg Transfer	0.72	0.76
Cutting	0.61	0.69
Clip Applying	0.60	0.63

Table 4.3: Correlation coefficients for the relationship between GOALS scores and predicted scores from the LSTM for every performance from each of the four BLUS tasks, including intermediate level performers, from Section 3.1.

there is a strong positive correlation, which has a Spearman coefficient of 0.89 and 91.67% accuracy.

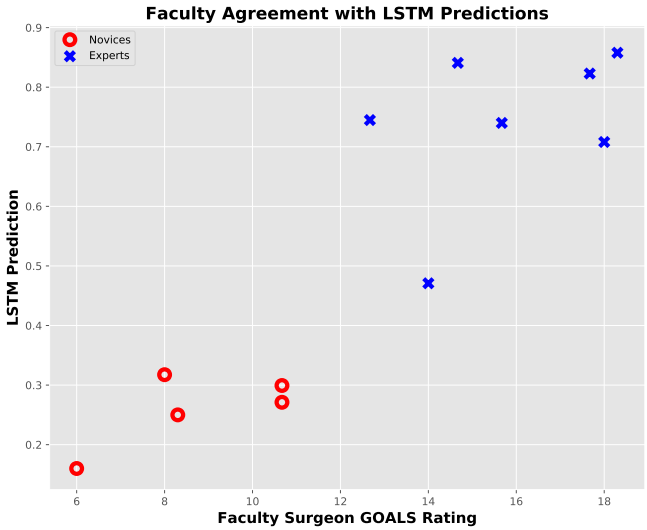


Figure 4.4: The LSTM predictions trained on crowd scores, for performances of suturing tasks as rated by faculty surgeons, which has a Spearman correlation of 0.89, from Section 3.1.

Once the model is trained with this kinematic data, it doesn't only allow classification of experts and novices, but in addition, allows the output to show the model's perceived skill of the surgeon through time for each of the subsequent segments, on a range of 0 to 1. One may view a video of a peg transfer task, and view how the LSTM model perceives the score of the surgeon as time progresses. It can be seen in one of these examples, that as the user makes mistakes, the perceived score decreases, and once they fix their mistakes and speed up, the LSTM is able to perceive their score increasing. A single frame example of this is shown in Figure 4.5.

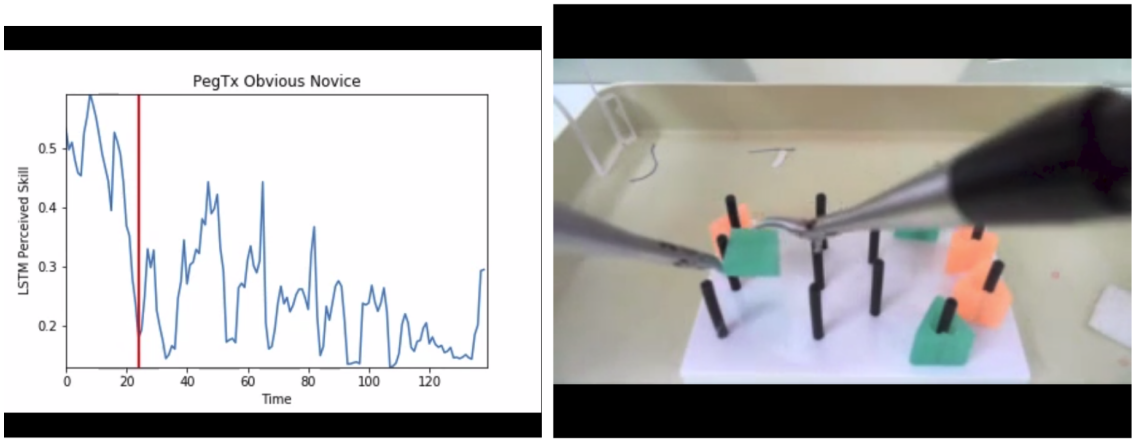


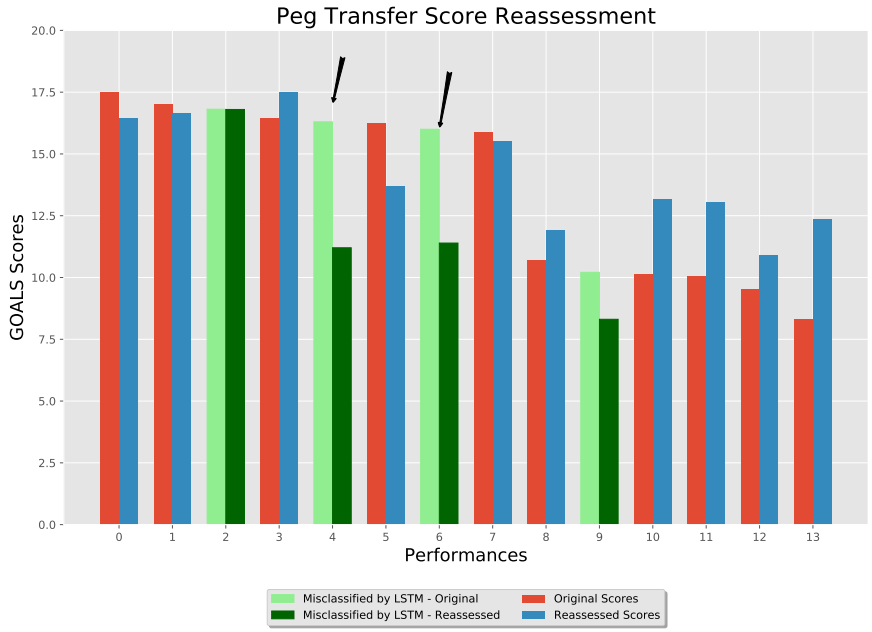
Figure 4.5: The perceived score of the LSTM on the left, with a video on the right of the same task.

4.1.4.2 Crowd Reassessment

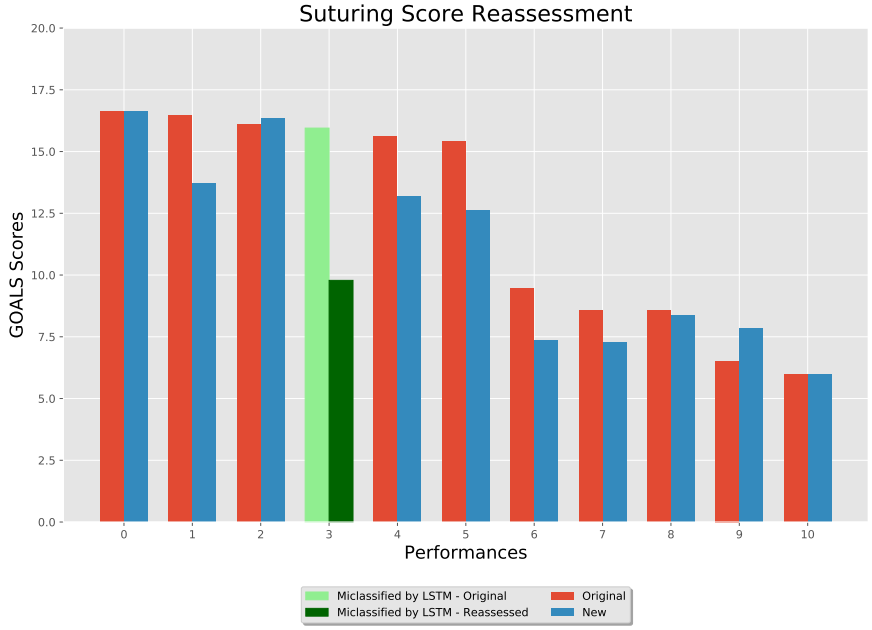
The suturing and peg transfer tasks were the two tasks with both high levels of classification accuracy as well as having a stronger correlation between scores and prediction levels, with a total of five misclassifications among the two tasks. The performances which were misclassified by the algorithm, which was trained on previously obtained crowd scores, were re-assessed by crowds. In addition to those performances,

5 randomly selected obvious novices and 5 randomly selected obvious experts were additionally chosen for reassessment.

Surprisingly, the range in the reassessed scores was quite lower than in the original assessments. This could have been caused from a variety of reasons, such as the newer user interface used as well as having had each video individually evaluated separately compared to the previous method in which videos were evaluated in batches. However, the reassessed ratings do still have a general agreement in ranking of performances as the original scores. Interestingly, if the new evaluations are normalized to be in the range of the old scores (5.96-16.61 for suturing and 8.31-17.5 for peg transfer), 3 of the 5 performances which were misclassified by the LSTM were reassessed to no longer be classified as having the skill level initially given to that performance, as shown in Figures 4.6 and 4.7. Figure 4.7 illustrates the propensity of the reassessed crowd scores to agree with previously obtained scores or to agree with the LSTM, based on the skill level predicted by the algorithm, and whether there was a misclassification. These figures suggest that the LSTM's classifications for those 3 performances could have been more accurate than the original ratings given to those performers.



(a) GOALS scores of misclassified peg transfer tasks before and after reassessment, as well as 5 obvious experts and 5 obvious novices, chosen randomly. Two of the 4 misclassified performances were reassessed to be more in line with what the LSTM predicted, indicated by the arrows.



(b) GOALS scores of misclassified suturing tasks before and after reassessment, as well as 5 obvious experts and 5 obvious novices, chosen randomly. The only misclassified performance was reassessed to be more in line with what the LSTM predicted.

Figure 4.6: Reassessment of misclassified suturing and peg transfer performances suggest crowds agree more with LSTM than previous crowd ratings, from Section 3.2.

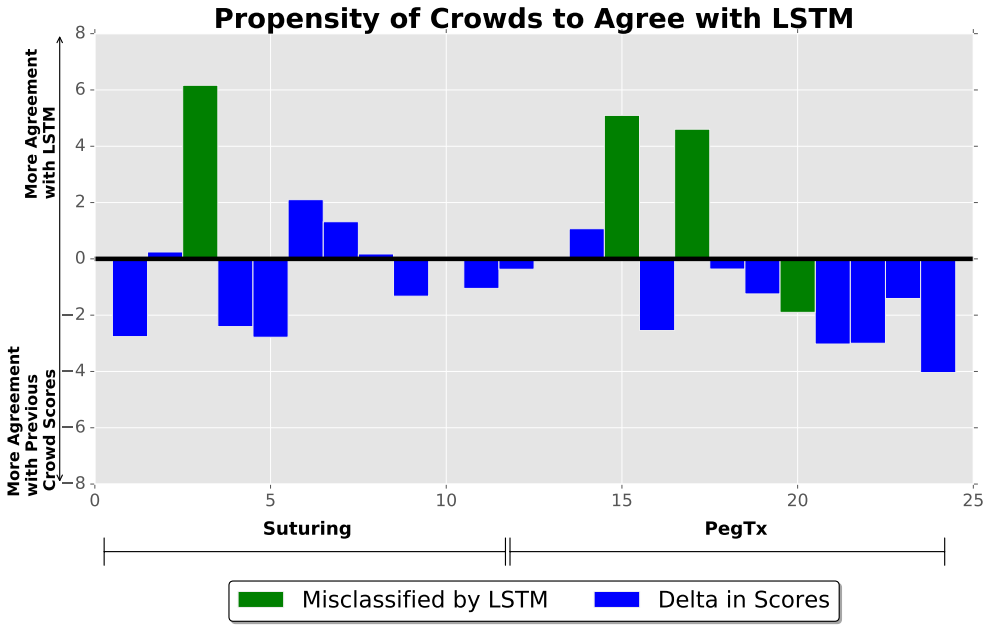


Figure 4.7: A bar plot of the magnitude of attenuation of the crowds to agree more with the LSTM on score reassessment, than with the original scores obtained.

4.1.5 Conclusion

The results from the bidirectional LSTM (Section 3.1), provide evidence in support of our hypothesis, that novice and expert surgeons do not exhibit expert performance metrics continuously throughout a task (and similarly for obvious novices). This suggests that temporal segmentation of kinematic tool motion analysis could provide more informative feedback of the skill of a surgeon, as compared to static summary performance metrics. The results, which also show good correlations between predicted score and actual score for several intermediate level performers, provide some additional evidence against overfitting, since these tasks were held out during the training phase.

Other popular methods of evaluating surgical skill such as computer vision algorithms which train on the frames of video, or popular automated performance metrics [41] could possibly be combined with this technique to create an even better classification model. This would further alleviate concerns about tool motion alone lacking important context that is still present in the video. Future iterations could enable giving correct predicted ratings of a performance, even during the surgical task performance, leading to near real-time skill feedback.

This study included some limitations. First and foremost, these tasks are simulated procedures, and the proposed algorithms and techniques may perform differently on real surgeries. We intend to test these hypotheses in the future on robot-assisted surgical data obtained from practicing surgeons. The proposed method also only analyzes tool motion data, which may not contain sufficient data required for complete skill classification [80]. The need for the reassessed scores to be normalized in order to have a comparison of crowd scores could be due to the techniques used in obtaining crowd evaluations which were different than the original evaluation methods. The authors acknowledge the results from the crowd reassessment could be in part due to differences in the user interface of the evaluation webpages used for the two different occurrences of testing.

4.2 Classification Using Video Data

Each of the videos from the Peg Transfer task were split into frames, and Shi-Tomasi corner detection points were computed for each. From here, the histogram of oriented gradients was computed by finding the derivative in both the x-direction, y-direction, and the temporal-direction, to find how these points change over time and space to get a sense of the optical flow. These points were then used in a similar way to how Zia used them to evaluate surgical skill [42]. Two expert videos of these points were concatenated to make one matrix, and a K-Means Clustering algorithm was used to divide the points into 'motion classes' which well symbolize the motion classes for expert motion in a peg transfer task.

Different numbers of motion classes were used to find the optimal number of classes to discriminate skill levels in video. Figures 4.8 and 4.9 show the motion classes in a sample video of a novice and expert.

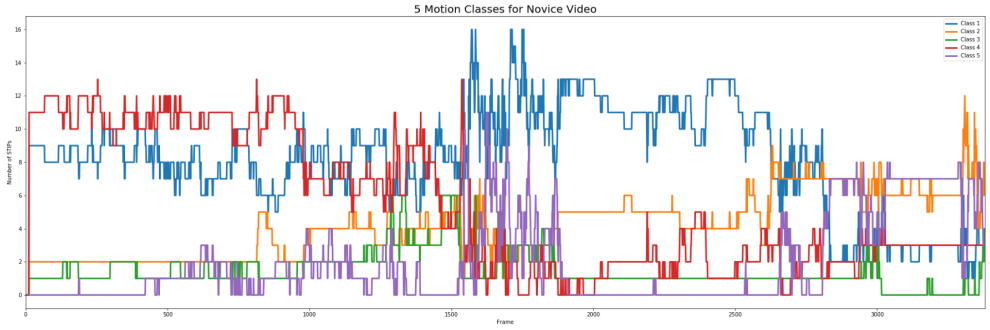


Figure 4.8: The frames of a novice video with the amount of points belonging to each motion class being shown.

These matrices can additionally be split into time windows which can be run through a Gaussian kernel function to visualize the similarity between frames in a certain time window. For example, in Fig. 4.10 we can see a frame kernel matrix for a portion of a novice video.

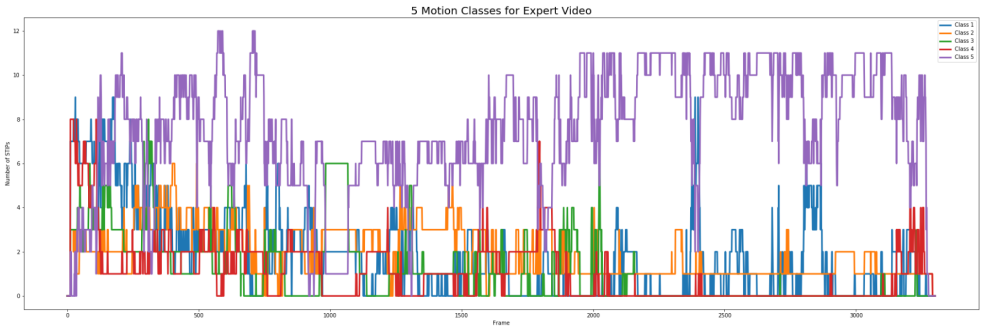


Figure 4.9: The frames of an expert video with the amount of points belonging to each motion class being shown.

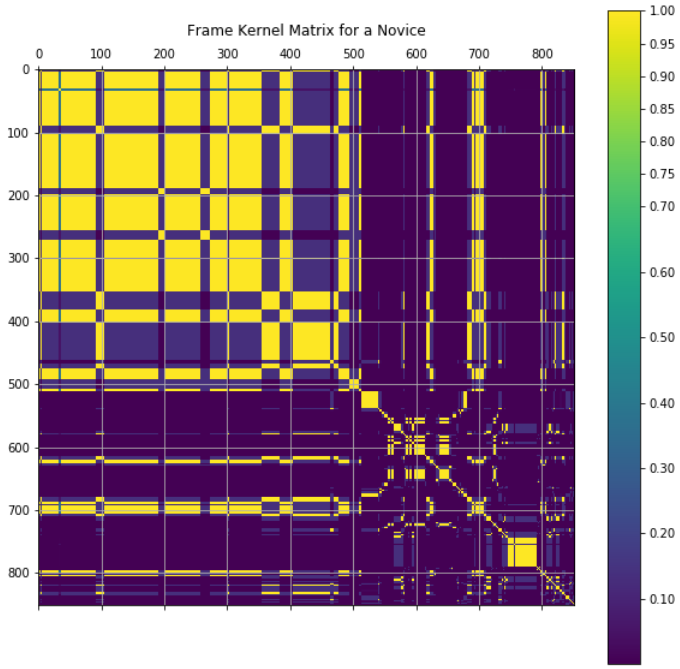


Figure 4.10: Frame kernel matrix for a portion of a time series, to visualize similarity between frames.

After obtaining the motion classes, the data was used as the input for the previously created bidirectional LSTM model, as well as a standard neural network model. Both of these models gave very poor results, with accuracy in classification reaching less than a coin toss. This proves to us that as it stands the amount of granularity in kinematic tooltip data is several degrees more complex than what is available in the video frames. The machine learning model is able to hone in on the complex details of the performance, the surgeon's movements, which most correlate to surgical skill. Using these algorithms with the frame-by-frame video data prevents the algorithm from being able to properly learn features when the algorithm is also being fed images which include much background information that isn't necessary. As it stands, it seems the best method of correctly classifying surgical technical skill from machine learning methods is by using tooltip kinematic data from which a model can most easily parse out the data. This method is able to currently outperform state-of-the-art machine learning methods with an accuracy of over 95% for tasks which closely replicate actual surgical performances.

Chapter 5

Speed Perception

This chapter discusses research regarding human speed perception and discusses experiments examining the existence of a bias in evaluating surgeons with different ratings when viewing videos of training exercises or surgeries at various speeds. The following is a reproduction of two separate publications: a dual journal and conference publicaton accepted to the International Journal of Computer Assisted Radiology and Surgery as well as the International Conference on Information Processing in Computer-Assisted Interventions at Munich, Germany in June 2020 [73], and a journal publication to the International Journal of Computer Assisted Radiology and Surgery, submitted in April 2020 [81]. Both of these publications contain their own abstract, introduction, methods, results, and conclusions, allowing each to stand alone.

5.1 The Effect of Video Playback Speed on Surgeon Technical Skill Perception [73]

5.1.1 Abstract

Purpose: Finding effective methods of discriminating surgeon technical skill has proven a complex problem to solve computationally. Previous research has shown that obtaining non-expert crowd evaluations of surgical performances is as accurate as the gold standard, expert surgeon review. The aim of this research is: (1) to learn whether crowd-sourced evaluators give higher ratings of technical skill to video of performances with increased playback speed, (2) its effect in discriminating skill levels, and (3) whether this increase is related to the evaluator consciously being aware that the video is manually manipulated.

Methods: A set of ten peg transfer videos (5 novices, 5 experts), were used to evaluate the perceived technical skill of the performers at each video playback speed used (0.4x-3.6x). Objective metrics used for measuring technical skill were also computed by manipulating the corresponding kinematic data of each performance. Two videos of an expert and novice performing dry lab laparoscopic trials of peg transfer tasks were used to obtain evaluations at each playback speed (0.2x-3.0x) of perception of whether a video is played at real-time playback speed or not.

Results: We found that while both novices and experts had increased perceived technical skill as the video playback was increased, the amount of increase was significantly greater for experts. Each increase in the playback speed by 0.4x was associated with, on average, a 0.72-point increase in the GOALS score (95% CI: 0.60-0.84 point increase; $p < 0.001$) for expert videos and only a 0.24-point increase in the GOALS score (95% CI: 0.13-0.36 point increase; $p < 0.001$) for novice videos.

Conclusion: Due to the differential increase in perceived technical skill due to increased playback speed for experts, the difference between novice and expert skill levels of surgical performances may be more easily discerned by manually increasing the video playback speed.

5.1.2 Introduction

Medical errors make up for a third of all deaths in the United States, one of the largest contributors of which are surgical errors [4]. Technical surgical skill is directly related to patient outcomes [72], but it remains a difficult computational task to correctly classify surgeons into skill levels with a compelling level of accuracy, i.e. never misclassifying an 'obvious novice' as an 'obvious expert' and vice versa - the MAC Criterion [70]. The de facto gold standard for evaluating technical skill is video evaluation by an expert surgeon using Likert-scale assessment metrics, in which evaluators submit ratings on an anchored scale of 1-5. Using crowds of *non-expert* evaluators are a surprisingly accurate way to inexpensively and rapdily obtain skill level ratings for videos of surgical performances, with a pass/fail rate capable of matching 100% of ratings by expert surgeons [22]. The fact remains, however, that humans can be biased in their thinking, and subjective metrics of rating performances can lead to results we would no expect from computational models of evaluation.

One of the most popular laparoscopic surgical skill assessment metrics is the Global Operative Assessment of Laparoscopic Skills (GOALS), which is the most common objective assessment tool for laparoscopic performance studies [24], [82]. The subdomains in this metric include: bimanual dexterity, tissue handling, efficiency, depth perception, and autonomy. Task time is not a direct metric used to evaluate the technical skill of laparoscopic surgeons with tools like GOALS, however time for task completion is often

seen as one of the most predictive objective forms of evaluating technical skill, as seen in Fig. 2.1.

Table 5.1: Likert-scale speed perception questionnaire for a manually sped-up video.

Score	**Does this video appear to have been altered to have an increased playback speed?**
(1)	Definitely sped-up video
(2)	Likely sped up video
(3)	Not sure / I don't know
(4)	Not likely sped up video
(5)	Definitely real-time playback speed (unaltered speed)

5.1.2.1 Biological Motion Perception

Previous research involving point light walkers (PLW), 10-12 dots of light illuminated on a screen to correspond to the human joints in walking motion, (illustrated in Fig. 2.13), revealed that humans are excellent at recognizing more subtle qualitative characteristics like the gender or emotion of a PLW by inferring the gait from a few moving points [62]. Once the PLW's walking gait speed is changed however, the ability to recognize gender is diminished [61]. This and other work suggest that biological motion perception is somehow tuned to the speed of the motion involved [2].

Due to task time being such a strong indicator of surgical technical skill, we speculate about whether a video artificially sped up to simulate a quicker time for task completion will lead to an altered perception of technical skill, or if human raters evaluate other, more nuanced, qualities of the performance.

5.1.2.2 Objective Metrics

Other metrics exist for measuring technical skill, with varying degrees of success at classifying novice and expert skill levels. These techniques utilize kinematic tool

tip data, as opposed to using video frames to make measurements. Metrics which use algorithms or computational methods to evaluate performance are sometimes referred to as Automated Performance Metrics (APMs) [41]. The simplest metric to measure, and often the most accurate, is time of task completion. The following APMs seek to compute information more complex than speed to account for technical skill, by conversely using the form of movements made.

Prior work has discovered that recovering stroke patients have a decrease in 'jerky' non-smooth movements as they progress in rehabilitation therapy [3]. The jerk cost of all tool movements performed in a task has also been used to measure accuracy and skill in areas in which skill is needed [79]. The jerk cost is computed by taking the integral of the squared jerk summed for both the left and right hand over the course of the performance, shown in Equation 5.1.2.2, in which T is total time of task performance and x is the magnitude of the movement.

$$\text{Jerk Cost} = \int_0^T |\dddot{x}(t)|^2 dt \tag{5.1}$$

Spectral Arc length (SAL) is also related to smoothness of a movement, and aims to use the tangential velocity of hand movements to compute the overall smoothness of a task. This has been used to differentiate skill levels of medical professionals in the past [83], [27]. To compute the SAL, the motion in a task must be segmented into specific hand grasping movements, with the corresponding speeds and durations known for each grasp. The Fourier magnitude spectrum transform of the speed profile is then computed and normalized with respect to its zero frequency value. The smoothness is then measured for each segmented grasp, shown in Equation 5.1.2.2, in which ω is

frequency, ω_c is cutoff frequency, and V is the Fourier magnitude spectrum of speed. Finally a weighted average is taken across all grasping motions to arrive at an SAL metric for the task as a whole.

$$SAL = -\int_0^{\omega_c} \left[\left(\frac{1}{\omega_c} \right)^2 + \left(\frac{d\hat{V}(\omega)}{d\omega} \right)^2 \right]^{\frac{1}{2}} d\omega,$$

$$\hat{V}(\omega) = \frac{V(\omega)}{V(0)} \tag{5.2}$$

Instead of relying on the form of a grasping movement, taking the sum of all movements which occur is an additional metric that can be used to evaluate skill in some domains in which it is believed that novice performers on average perform more grasping movements than experts. A movement can be calculated in several ways, but the most common way of recording is to define a threshold of speed at which when the performer's speed falls below that threshold, a movement has ended and a new grasping motion starts once the magnitude of the velocity has risen above that threshold again. Counting each time this threshold is passed gives this total count of all movements conducted during the task [84].

All of these APMs which may be used to compute the skill of a surgeon have in common that once the speed playback of the data is manipulated, the results normally do not lead to any more separation between skill levels, as changing the magnitude of speed will not affect calculations to a large degree.

Our motivation in this work is to learn whether, by increasing video playback speed, novices and experts will be perceived by crowds as more skilled when they appear to be moving faster, *and* if experts and novices will have different rates of perceived change as the playback speed is increased. We will do this by evaluating the ability to discriminate obvious novices from experts of both human raters and popular tool motion metrics

(APMs), at different playback speeds. We expect each APM's ability to discriminate skill to operate as the experimental control in this study, with a negligible change in separation between skill levels as speed is increased. Finally, we seek to investigate how an evaluator's likely conscious awareness of whether a video is sped up or slowed down relates to their perception of skill.

5.1.3 Methods

5.1.3.1 Dataset

This study used the Basic Laparascopic Urologic Study (BLUS) dataset, described in detail in [22]. This dataset arose from a gap in the field, in which no educational surgical certification process existed for urologic surgery, as opposed to how the Fundamentals of Laparosocopic Surgery (FLS) exists for general surgical procedures [75–77]. The BLUS training curriculum aimed to address urology appropriate skills improvement by recording video performances in an initial validation project of over 450 videos [78].

This dataset contains 454 videos of surgical performances consisting of four surgical tasks (110 peg transfer, 110 pattern cutting, 115 suturing, 119 clip applying), which are performed by medical students, urology residents, fellows and faculty surgeons from eight academic urology training centers in the United States [79]. Each trial of a surgeon performing one of the four tasks was recorded at 30 fps with a fixed camera-position of the laparoscopic tools interacting with the training field. Each trial additionally has kinematic data, sampled at 30 Hz, logging the tooltip positions, grasping force, and the jaw angles during the performance, as well as demographic information for each performer being obtained. A GOALS score for each performance was also previously obtained from either expert or crowd evaluation.

In previous research the Peg Transfer task has been shown to be one of the most

easily differentiable tasks for surgical technical skill. Although the Peg Transfer task is criticized as the least clinically relevant, its clear ability to separate novice/expert skill levels was preferred to explore the research questions herein. Ten videos from the peg transfer task were used, in which five 'obvious experts' and five 'obvious novices' were chosen as baseline definitions of skill [70]. Here an obvious novice is defined as someone who should never be allowed to operate and obvious experts as surgeons who should never be disqualified from operating. An obvious expert was chosen such that the performer was in the top 20% of experience levels (attending surgeon or faculty urologist), previously obtained GOALS scores, and task completion times of all peg transfer tasks. The obvious novices were chosen in the same fashion such that they were in the bottom quintile of these domains. This method aimed to provide two well discriminated clusters of skill levels.

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.10 to watch and evaluate a video, depending on the video duration. A web domain was created for which Turkers would be redirected to, where they submitted a consent form and were asked questions about videos. Two different kinds of experiments were conducted: *technical skill perception* and *speed perception*.

5.1.3.2 Experiment 1: Technical Skill Perception

Technical skill perception was measured by surveying non-expert crowds to give each video performance a GOALS score by rating each of the 4 subdomains shown in Table 2.1. Forty "turkers" were recruited per video, in which each video at each playback speed was independently submitted to the website in order to avoid a grouping bias. Videos were altered to speeds in the range 0.4x-3.6x, moving in intervals of 0.4, edited using FFmpeg [85], in which frames were either taken out or added in order to create

the resulting playback speed. The score from each of four subdomains (Depth Perception, Bimanual Dexterity, Efficiency, and Tissue Handling) were summed to create a cumulative score for each performance in the range of 4-20. The mean of each video's cumulative GOALS score was recorded, with a 95% confidence interval.

5.1.3.3 Experiment 2: Speed Perception

The video playback speed perception test was conducted by using one novice and one expert from the peg transfer task, and asking the evaluator if they thought the video was being manually edited. If we showed them a video with a decreased playback speed, the worker was asked whether they thought the video was running at real-time speed or was manually edited to be slowed down. In the same way, workers were asked if the video was being sped up or played at real-time playback speed for videos which had an increased video playback speed. These Likert-scale questions are shown in Table 5.1 for a video in which the playback speed was increased. Speeds tested ranged from 0.2x-3.0x, moving in intervals of 0.2x to verify the range of speeds to use for the technical skill perception tests.

5.1.3.4 APM Validation at Each Playback Speed

The Automated performance metrics mentioned in Section 5.1.2.2 were also recorded at these various playback speeds for comparison to crowd evaluations. These were re-calculated by scaling the magnitude of the speeds in the kinematic data to be equivalent to the speed displayed in a given video playback speed, then computing these objective metrics. For the jerk cost, this involved using a Holoborodko smooth noise-robust differentiator [86]. For the others this involved simply scaling the magnitude of speed and inputting the result to the pre-created function for each corresponding APM. All APMs were computed in Python and MATLAB, [87, 88].

5.1.4 Results

5.1.4.1 Technical Skill Perception

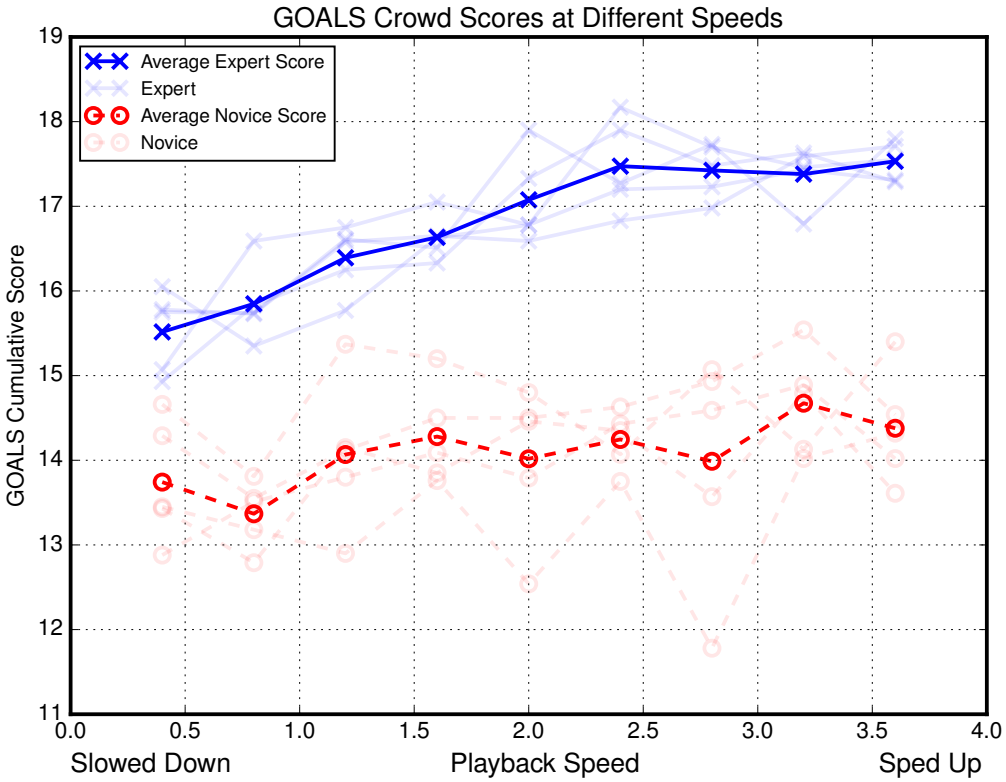


Figure 5.1: All mean crowd evaluations from each novice and expert peg transfer task at various video playback speeds. (Each solid marker indicates N = 40).

The mean of the GOALS score for each obvious expert and novice are shown in Fig. 5.11. For expert videos, each increase in the playback speed by 0.4x was associated with, on average, a 0.72-point increase in the GOALS score (95% CI: 0.60-0.84 point increase; $p < 0.001$). On average these scores appear to increase within a sublevel of the playback speeds around 0.5x to 2.4x, and then level out at all remaining playback speeds. For novice videos, each increase in the playback speed by 0.4x was associated with, on average, a 0.24-point increase in the GOALS score (95% CI: 0.13-0.36 point

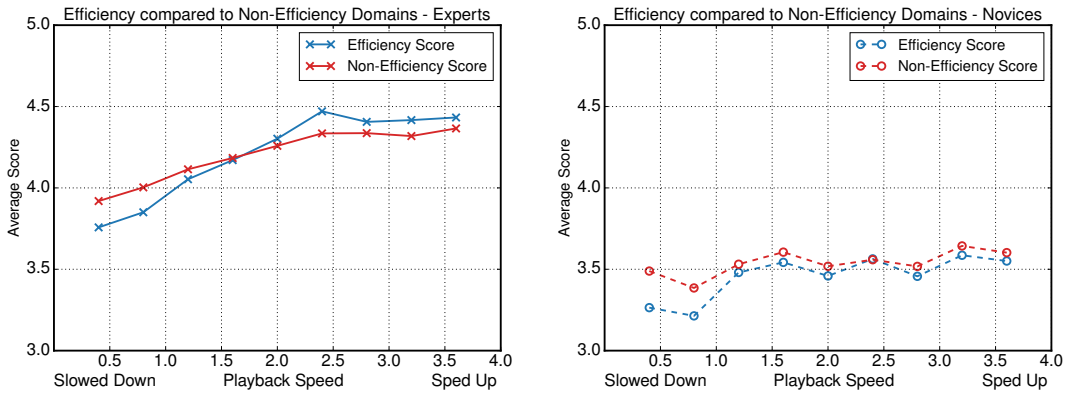


Figure 5.2: The efficiency subdomain as compared to the average of the other GOALS subdomains, for experts and novices.

increase; $p < 0.001$). Thus, while both experts and novices had increased perceived technical skill as the playback speed was increased, the gain was significantly greater for experts. The experts had, on average, a 0.47-point greater increase (95% CI: 0.31-0.64; $p < 0.001$) in the GOALS score, compared to novices, for each 0.4x increase in playback speed. Fig. 5.12 shows the increase in the efficiency subdomain as well as the average of the other three domains to visualize whether efficiency (seen as the most related to speed) is the only increasing domain. As shown, it is clear that the other domains increased at nearly the same rate.

The APMs which were calculated with the manipulated kinematic data are shown in Fig. 5.3. As shown, most metrics show very little change at each of the different speeds, but even if they change, performers' technical skills are not able to be discriminated between expert and novice skill any more easily. These metrics were also calculated by incorporating a combination of scaling the speeds and removing indices corresponding to the frames which would be removed from videos when the playback speed is increased. These figures aren't included, but the results exhibit the same lack of separation between the skill groups at various playback speeds. Fig. 5.4 shows the difference in mean for

experts and novices, tested with both the crowd scores and the highest performing APM, spectral arc length, to illustrate the difference in skill discrimination between crowd workers and the most accurate APM as speed is increased.

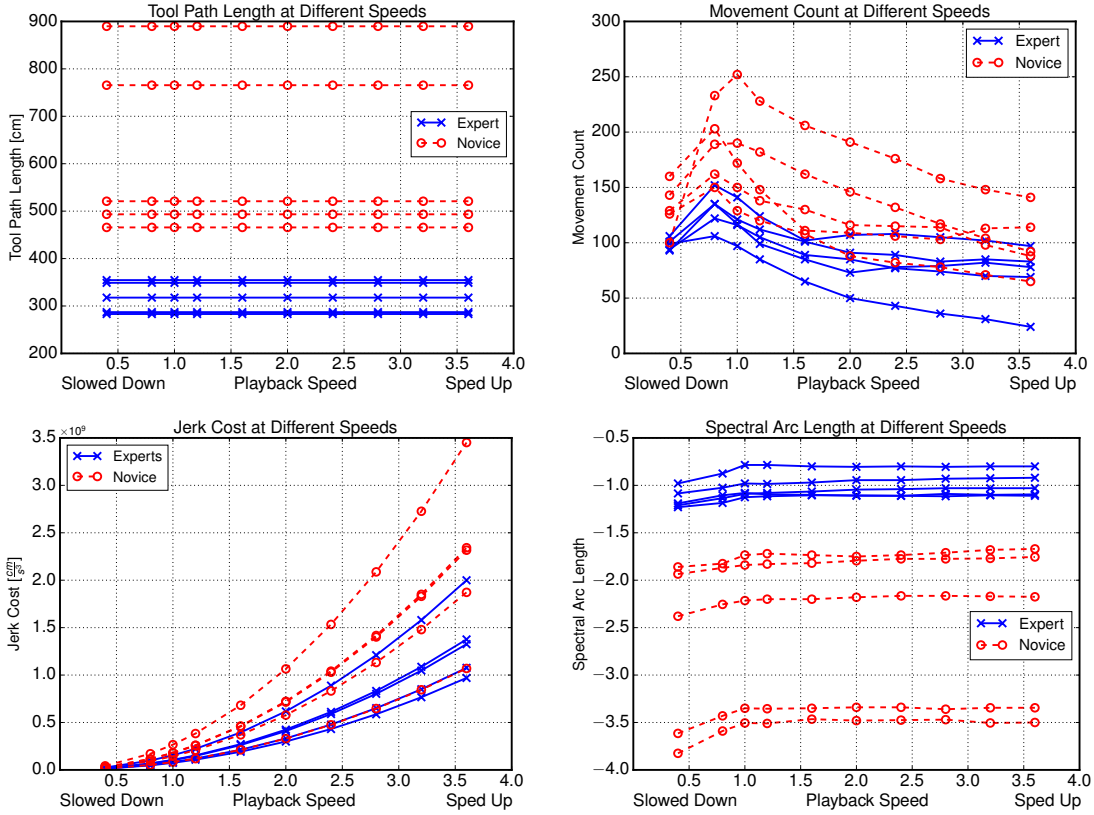


Figure 5.3: Objective technical skill metrics for the obvious experts and novices across various speeds.

5.1.4.2 Playback Speed Perception

The speed perception results for an obvious novice and obvious expert are plotted in Fig. 5.5a. The scores were obtained by treating each of the scores in Table 5.1 as a score, and obtaining the mean score for each video evaluated, such that a higher score leads to a higher perception of the video being played at a real-time playback speed. Both

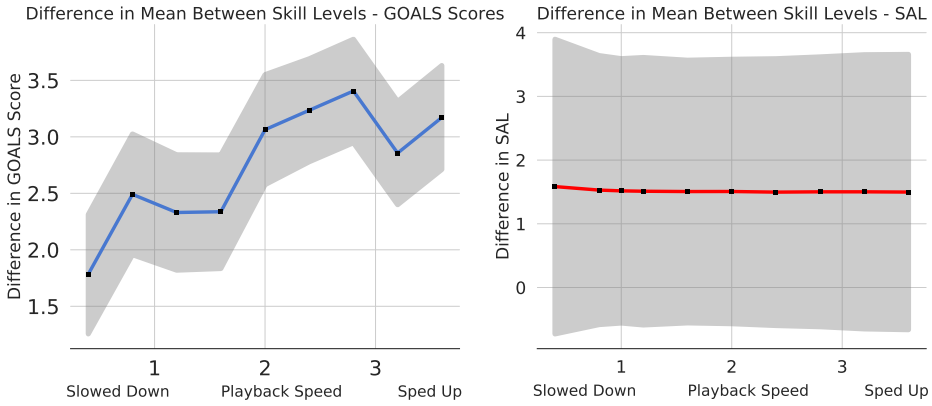
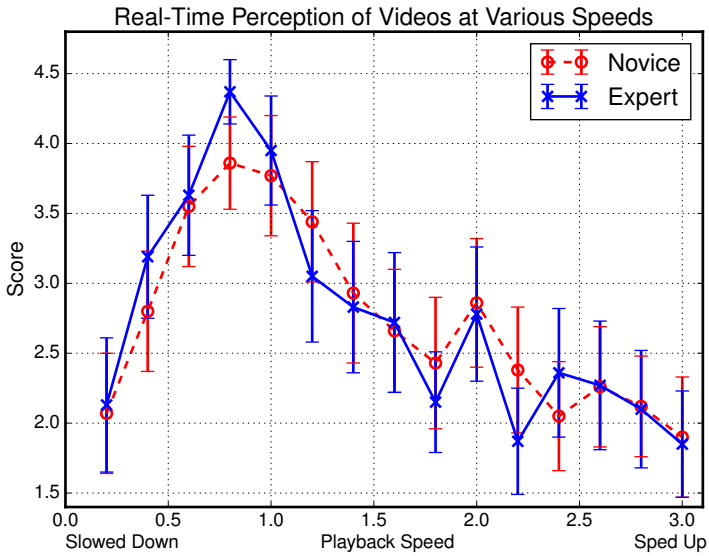


Figure 5.4: Difference in mean between skill levels from the highest performing APM and from crowd scores, with a 95% CI in the shaded regions.

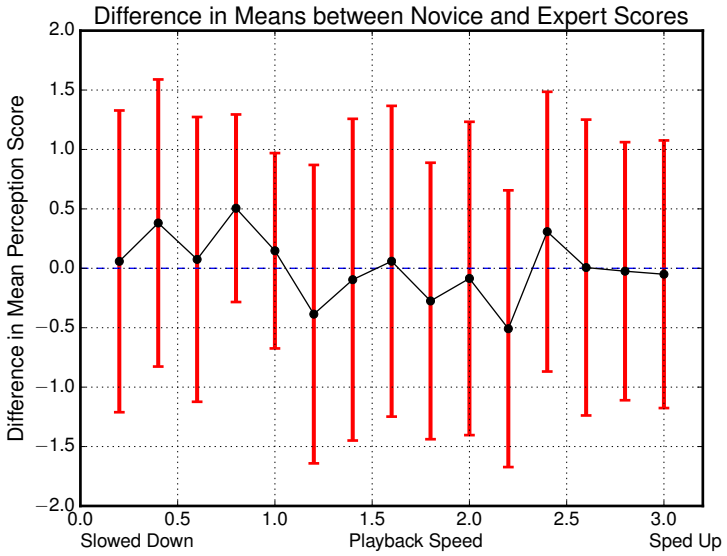
novices and experts displayed similar perceptions of the video either having an increased or decreased video playback speed. In Fig. 5.5b, a two sample independent t-test was computed between the two groups at each of the playback speeds, to visualize the difference in mean speed perception between experts and novices. All of the speeds show no significant difference in playback speed recognition. The perceived video playback speed scores are compared with the same video's crowd scores for the GOALS assessment metric in Fig. 5.6.

5.1.5 Conclusion

The results from the technical skill perception study give support to our initial hypothesis that increasing the video playback speed would increase the ratings of surgical performances. Surprisingly, however, we discovered that novice performances receive a much lower increase in score, which is almost negligible. This finding elucidates the notion that despite increasing the video playback speed of a novice performance, non-expert crowd workers are still able to spot the more obvious mistakes made by these

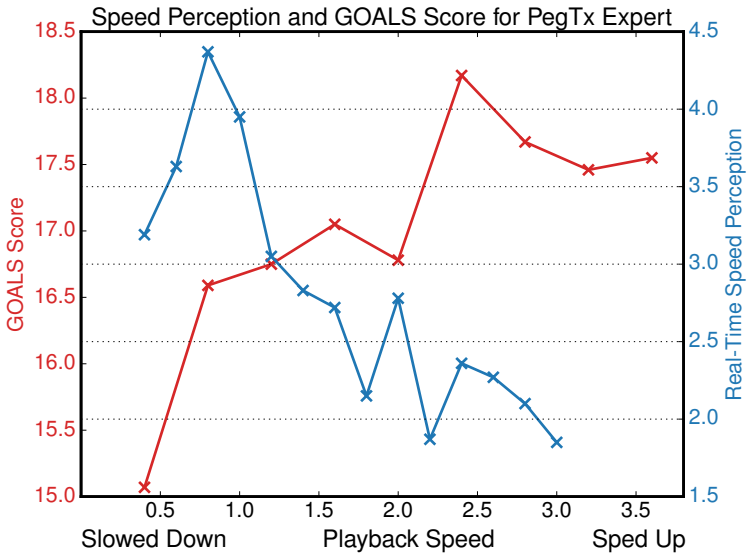


(a) All mean crowd evaluations of video speed perception at various video playback speeds. (Each solid marker indicates N = 40).

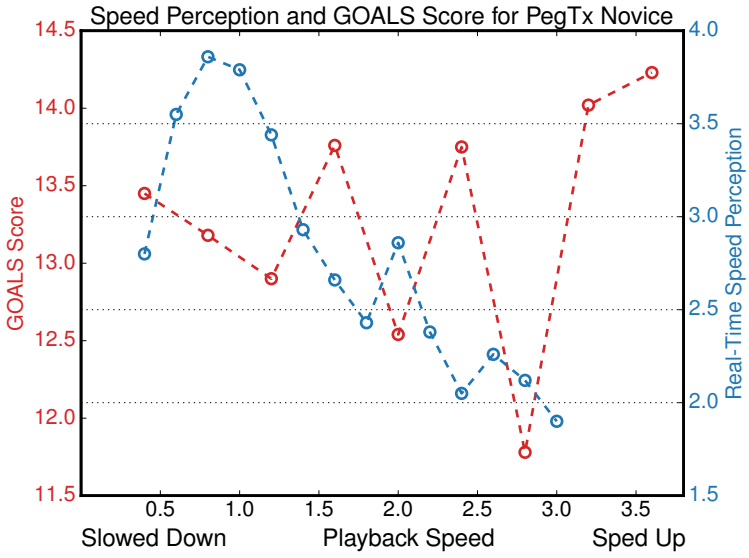


(b) Difference in Mean between Expert and Novice evaluations.

Figure 5.5: Comparison of Novice and Expert Speed Perception at each playback speed. (There were $N = 40$ *unique human evaluations/skill level/playback speed*. The error bars represent 95% confidence intervals.)



(a) Expert



(b) Novice

Figure 5.6: Speed Perception compared with technical skill perceived at various speeds.

novices. However, it appears that crowd evaluators are biased to speed when they evaluate expert performances. This could be due to expert performers appearing as though

their movements are even smoother at quicker playback speeds and the few mistakes they are making being 'washed out' or not emphasized at quicker playback speeds. With 135,000 surgeons in the U.S. and the growing need to objectively quantify surgical skills to ensure public safety, methods to rapidly triage technical skill are required. Our observations may provide an easier and cheaper way of discriminating novices from experts than with using expert surgeon evaluation.

To illustrate whether the 'Efficiency' domain of the GOALS assessments was the only domain increasing, as this domain is most closely related to speed, a scatter plot for the average efficiency domain and average of the other three domains for novices and experts is shown in Fig. 5.12. The novices clearly show no difference in Efficiency vs. the other three domains. In addition, the three remaining domains for experts appear to increase at almost the same rate as the Efficiency domain. It appears that crowds are also biased to give higher ratings in other domains which aren't necessarily related to speed.

The results from the real-time playback perception study validate our initial hypothesis by showing us that crowds are able to accurately discern when a video has been manually edited to have a different playback speed. Although, as shown in Fig. 5.5a, there is apparently no difference in perception between an obvious novice performance and an obvious expert performance.

As shown in our comparison to objective metrics, no other used method of objectively obtaining technical skills scores will show the same amount of improvement for increases in playback speed. There may be information that humans can decipher which objective metrics or machine learning algorithms cannot.

We conclude that increasing the video playback speed of performances of dry lab laparoscopic training tasks could provide a cheap and easy way to discriminate experts from novices as the separation in GOALS scores between novices and experts appears

to increase as video playback is increased. A limitation of this study is that we only sampled ten videos of laparoscopic training procedures. Additional investigation with a larger dataset of more clinically relevant performances is required to conclude whether this observation extends to actual surgical case evaluation.

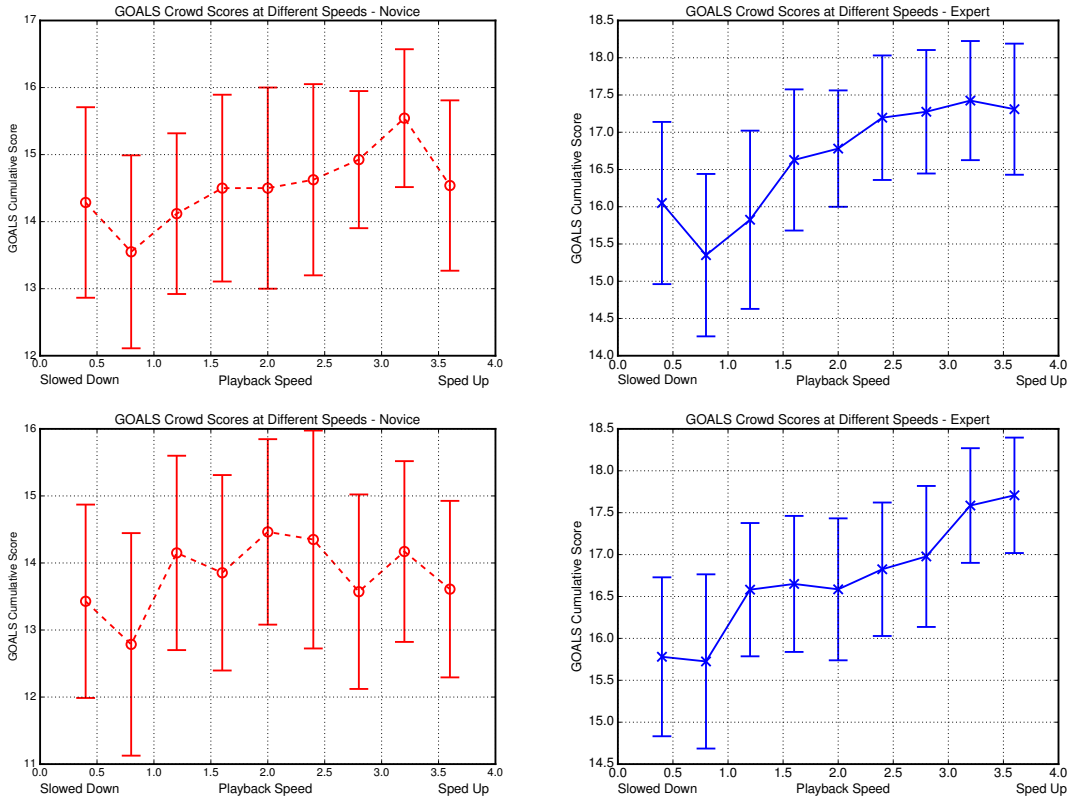


Figure 5.7: Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is $N = 40$, with 3600 total evaluations.

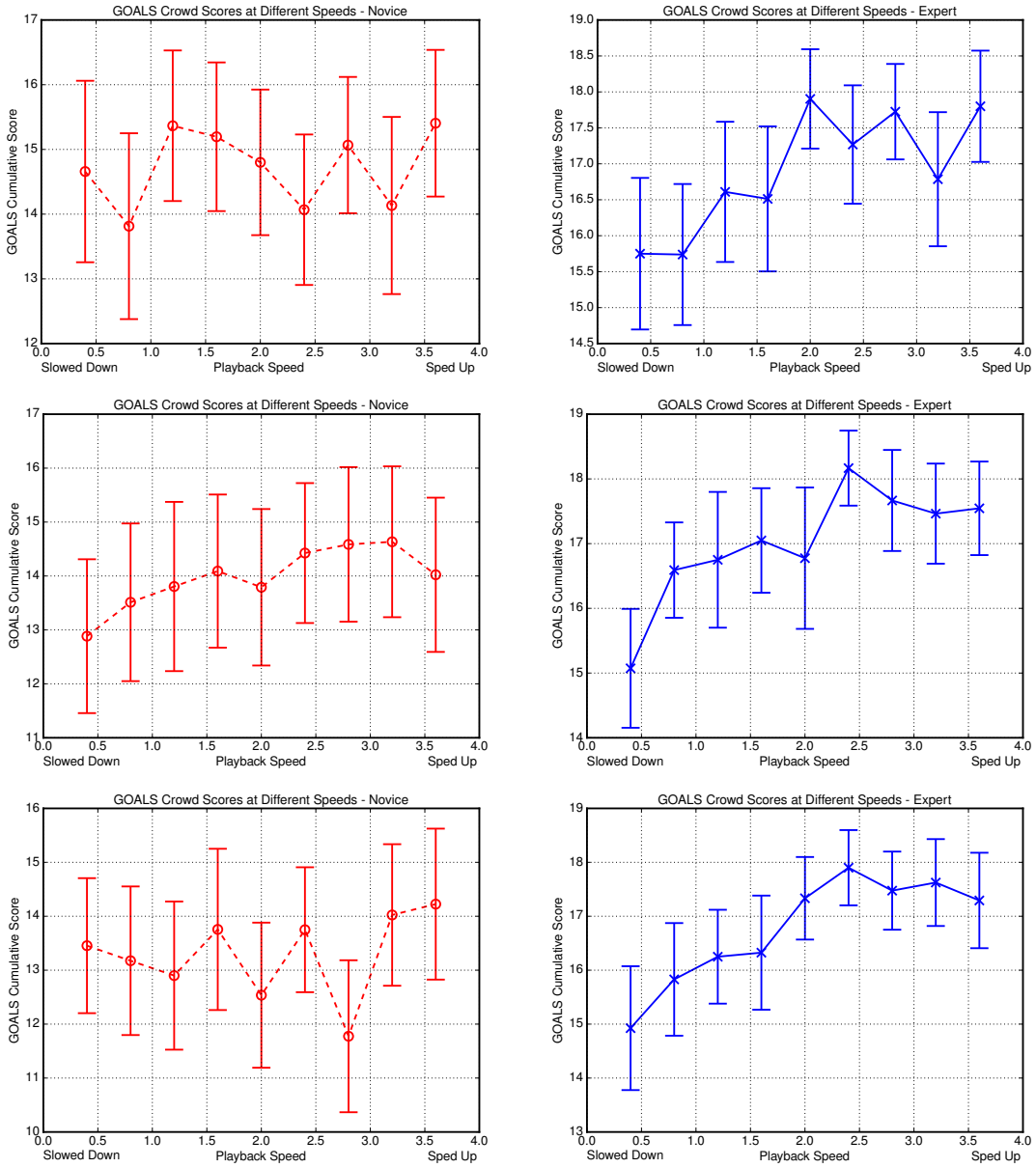


Figure 5.8: Crowd GOALS evaluations for Experts and Novices at various playback speeds. Each point is $N = 40$, with 3600 total evaluations.

5.2 The Effect of Video Playback Speed on Perception of Technical Skill in Robotic Surgery [81]

The previous study had two major shortcomings, it was a small amount of data and the videos being evaluated weren't from actual surgical footage. This following study aimed to dive deeper into the notion of speed perception in surgical settings and learn the effects of studying a smaller subset of skill, surgeons who are all above average performers.

5.2.1 Abstract

Purpose: Finding effective methods to evaluate surgeon technical skill has proven a complex problem to solve computationally. Previous research has shown that obtaining non-expert crowd evaluations of surgical performances concords with the gold standard of expert surgeon review, and that faster playback speed increases ratings for videos of higher-skilled surgeons in laparoscopic simulation [22], [73]. The aim of this research is to extend this investigation to real surgeries that use non-expert crowd evaluations. We address two questions (1) whether crowds award more favorable ratings to videos shown at increased playback speeds, and (2) if crowd evaluations of the first minute of a surgical procedure differ from crowd evaluations of the entire performance.

Methods: A set of 56 videos of practicing (non-novice) surgeons including robotic prostatectomy, hysterectomy, and partial nephrectomy (for 28 "expert" surgeons, and 28 who are "proficient"), were used to evaluate the perceived technical skill of the surgeons at each video playback speed used (0.4x, 1.2x, 2.0x, 2.8x, and 3.6x) for the first minute of the previously rated performance, using the Global Evaluative Assessment of Robotic Skills (GEARS) assessment criteria. Each video was subsequently rated at 1x

speed to obtain objective ratings for the first minute of the surgical procedure.

Results: Crowds on average did rate videos higher as playback speed was increased. This effect was observed for both proficient and expert surgeons. Each increase in the playback speed by 0.8x was associated with, on average, a 0.16-point increase in the GEARS score for expert surgeons and a 0.27-point increase in GEARS score for proficient surgeons, with both groups being perceived as obtaining relatively equal skill at the fastest playback speed. It was also found that 22 out of the 56 surgeons were perceived to be significantly different in skill when just viewing the first minute of performance, with 11 of the 28 surgeons in both skill categories being rated as belonging to the opposing category.

Conclusion: The observed increase in skill ratings with video playback speed replicates findings for laparoscopic experts in [73], and extends to the context of real robotic surgeries. The change in perceived technical skill due to increased playback speed for experts and proficient surgeons suggests that crowds do seem biased in rating surgeons as more highly skilled when they appear quicker, even if speeds seems unrealistic. Furthermore, the large differences in skill labels when comparing the first minute of surgery to the entire 15 minute video warrants further investigation into how much perceived skill ratings vary in time (sub-task level) vs. summative metrics (task level).

5.2.2 Introduction

A third of all deaths in the United States are caused by medical errors, and surgical errors are one of the largest contributors to this [4]. Technical surgical skill is directly related to patient outcomes [72], but it remains a difficult computational task to correctly classify surgeons into skill levels with a compelling level of accuracy, i.e. never misclassifying an 'obvious novice' as an 'obvious expert' and vice versa - the MAC

Criterion [70]. The de facto gold standard for evaluating technical skill is video evaluation by an expert surgeon using Likert-scale assessment metrics, in which evaluators submit ratings on an anchored scale of 1-5. Using crowds of *non-expert* evaluators are a surprisingly accurate way to inexpensively and rapidly obtain skill level ratings for videos of surgical performances, with a pass/fail rating that matches 100% of pass/fail ratings by expert surgeons [22]. The fact remains, however, that humans can be biased in their thinking, and subjective metrics of rating performances can lead to results we would not expect from computational models of evaluation.

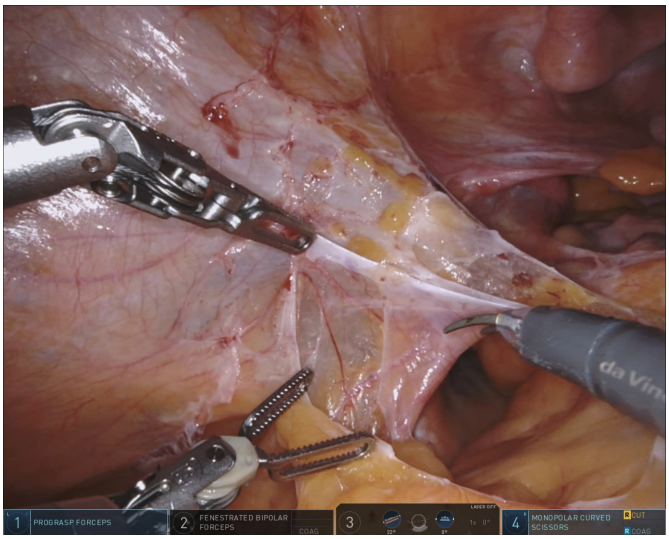


Figure 5.9: Frame from a video collected for this study, recorded using a daVinci surgical robot, Intuitive Surgical (Sunnyvale, CA).

A popular laparoscopic surgical skill assessment metric is the Global Evaluative Assessment of Robotic Skills (GEARS), which is the most common assessment tool for robotic surgery skills [89]. The subdomains in this metric include: bimanual dexterity, efficiency, depth perception, force sensitivity and robotic control. Task time is not a direct metric used to evaluate the technical skill of laparoscopic surgeons with tools like GEARS, however time for task completion is often seen as one of the most predictive

objective forms of evaluating technical skill [13], [41]. However, there must be ways of objectively evaluating skill between multiple performances which were completed in the same time span.

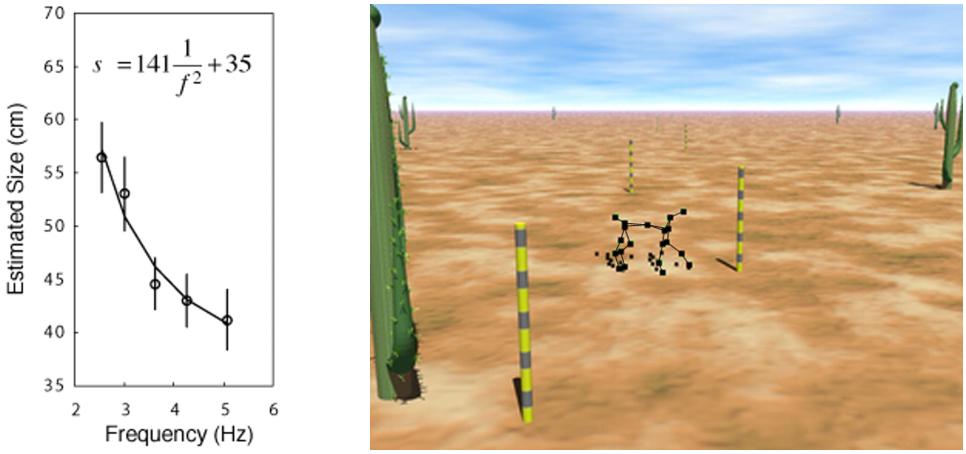


Figure 5.10: Illustration of research results (left) which found the frequency of gait in simulated animals (right) affects the estimated size [**?**].

5.2.2.1 Changes in Perception due to Speed

Research in biological motion processing has observed a relationship between the gait frequency and estimated size of animals and other objects which replicate the joints of the human body. It was found that when a computer simulation was used to artificially modify the gait speed of these objects to unnaturally high speeds, study participants perceived the size of the moving objects as changing. As the gait frequency was increased, the estimated size of the object decreased, shown in Fig. 5.10, [66], [61]. The results from this study may suggest a link between how humans evaluate biological motion and the speed of a movement. It is possible this phenomenon extends to other areas of evaluation, such as technical skill in surgery.

5.2.2.2 Technical Skill Ranges

Previous work has shown that crowds are able to discriminate the levels of novice and expert surgeons and their scores change as playback speed changes - even into motions appearing artificially sped up [73]. This was a limited study using a small dataset of laparoscopic training exercises from the Basic Laparoscopic Urologic Skills dataset [78], [79], with extremes of skill levels (e.g. "obvious novice" medical students with no suturing experience). To test whether this phenomenon relates to actual surgical procedures, a larger dataset must be used which incorporates videos of real surgical footage and more practical ranges of surgeon skill. However, surgeons which are allowed to safely operate on patients can't ethically be "obvious" novice surgeons, according to the MAC Criterion [70]. This means the surgeons in a dataset of real surgical footage will comprise a smaller range of technical skill levels, with all surgeons being more comparably rated. This necessitates the need to further separate experts into two groups: proficient and expert surgeons. Proficient surgeons as stated in [90] are defined as a surgeon who is well advanced in any branch of knowledge or skill. We will adopt this as our term for surgeons who are in the bottom quintile of scores from our surgical videos, as they are still well above average skilled surgeons.

Our motivation in this work is to investigate whether, by increasing video playback speed in real robotic surgery, surgeons will be perceived by crowds as more skilled when they appear to be moving faster, *and* how these effects vary from expert and proficient surgeons. We will do this by evaluating the ability of non-expert crowd workers to discriminate proficient surgeons from experts. Finally we seek to examine the effect of video duration on ratings following [91].

5.2.3 Methods

5.2.3.1 Dataset

This study used a novel Robotic Surgery Readiness (RSR) Study dataset, which consists of 343 videos of live robotic surgeries, with matching kinematic data also recorded, though unused here. These surgeries were performed by attending surgeons and trainees in urology, gynecology, and general surgery at the University of Washington Medical Center and the Puget Sound Veterans Administration. Each video was manually edited to include roughly the first 15 minutes of surgical activity performed by the surgeon. There are a wide variety of surgical procedures recorded from this dataset including: prostatectomy, cystectomy, hysterectomy, partial nephrectomy, and sacrocolpopexy. An image of a frame from one of these surgeries is shown in Fig. 5.9. A GEARS score for each performance was also obtained from crowd evaluation.

The range of scores for the RSR videos was fairly small, with most lying between 20-22 out of 25. In an effort to get the largest possible range of skill from this dataset, 28 performances from the top quintile of scored performances and 28 from the bottom quintile of performances were used and given labels of 'expert' and 'proficient', respectively, keeping in mind that almost all of the performances would have objectively been considered expert-like. For semantic analysis, the first minute of surgical activity was extracted from each 12-15 minute video for analysis by crowds.

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.40 to watch and evaluate a video. A web domain was created for which Turkers would be redirected to, where they submitted a consent form and were asked questions about videos. Two different kinds of experiments were conducted: *technical skill perception at different playback speeds* and *sub-task level skill labeling.*

5.2.3.2 Technical Skills Perception at Different Playback Speeds

Technical skill perception was measured by surveying non-expert crowds to give each video performance a GEARS score by rating each of the 5 subdomains shown in Table 2.2. Forty "turkers" were recruited independently for each video, in which each video at each playback speed was independently submitted to the website in order to avoid a grouping bias. Videos were altered to play at 0.4x, 1.2x, 2.0x, 2.8x, and 3.6x, edited using FFmpeg [85], in which frames were either taken out or added in order to create the resulting playback speed. The score from each of five subdomains (Depth Perception, Bimanual Dexterity, Efficiency, Force Sensitivity, and Robotic Control) were summed to create a cumulative score for each performance in the range of 5-25.

A linear mixed effects model was used to analyze the significance of the various speeds to the evaluations received. The significance of the time spent on reviewing a video and the labels given to the surgeons were also examined. It was hypothesized that each Mechanical Turker on the site should be assumed to have a different slope, to match the difference with which they relatively evaluate different videos. The evalutor ID given to each rater was assigned a random effect. In addition, fixed effects for both the speed at which the video was played and the amount of time spent on the video by the evaluator were controlled for in the model. The mixed model was compared with a null model, which did not include the fixed effects, using ANOVA hypothesis testing and comparing the relative information criteria and correlation between the scores and the various parameters. All data aggregation and statistics were calculated in Python 3.6 [87] and R [92].

5.2.3.3 Sub-Task Level Skill Labeling

To learn if the first minute of surgical activity displayed the same level of technical skill as the entire 15 minute video and should have the same label, the first minute of each video at 1x (normal playback speed) was submitted to crowds for review. The mean of these new scores were then calculated, and the 50th percentile was used as the cut-off point to signify in this dataset of 56 videos, whether a surgeon would be labeled as a proficient or expert level surgeon. The differences in these labels were then analyzed further and visualized, to find if the previously obtained scores for the entire 15 minute video possessed significantly different skill levels in the first minute, such that the label of the corresponding surgeon was different than what we had originally obtained.

5.2.4 Results

Table 5.2: Results from the linear mixed effects model testing speed and time spent on evaluation

Fixed Effects	**Estimate**	**Standard Error**	**df**	**t value**	**Pr (> | t |)**
Initial linear mixed effects model (BIC = 52465.36)					
Speed	0.046	0.013	13360	3.47	5.31e-04
Elapsed Time	-0.0023	3.47e-4	13720	-6.57	5.18e-11
Final linear mixed effects model (BIC = 52410.60)					

5.2.5 Data Demographics

Table 5.3 summarizes the types of robotic surgery and the demographic data from the $N = 56$ videos (and surgeons) used in this work.

Table 5.3: The types of robotic surgery performed and demographics of the surgeons from the 56 videos used.

Surgery Type	N
Prostatectomy	17
Hysterectomy	11
Nephrectomy	5
Cystectomy	4
Other	19

Demographic	N
Male	17
Female	11
Mean Age	46.70
Mean Yrs. Experience	12.85

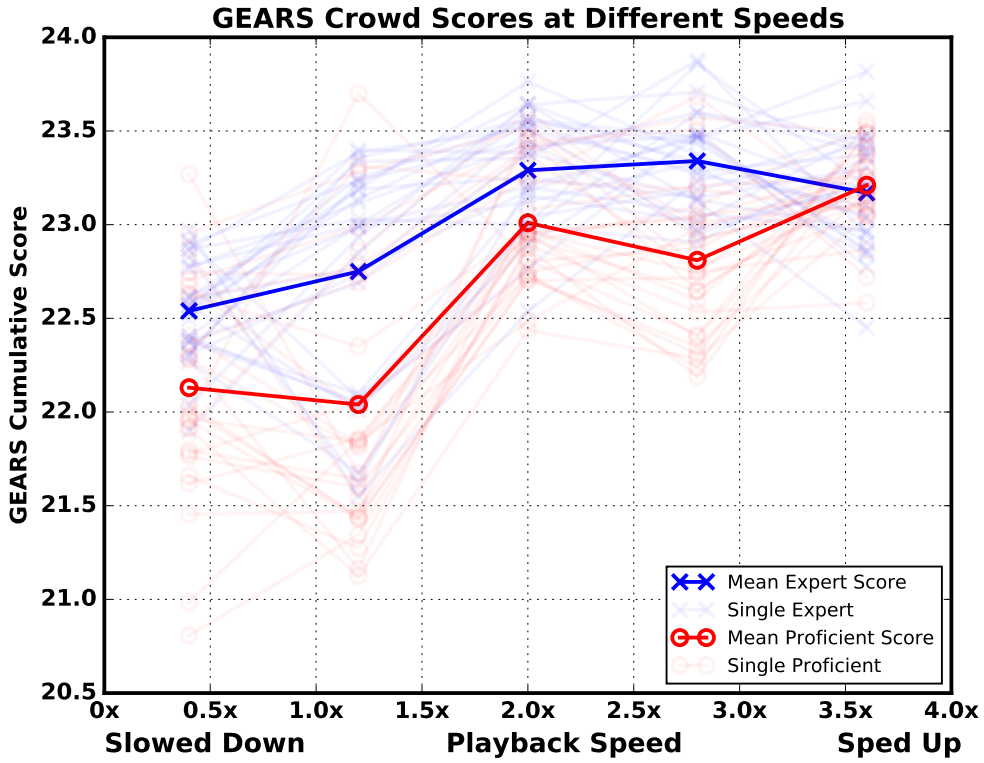


Figure 5.11: All mean crowd evaluations (bold lines) from each proficient and expert surgeon at various video playback speeds. Each single surgeons video (semi-transparent colors) indicates ratings from N = 40 turkers.

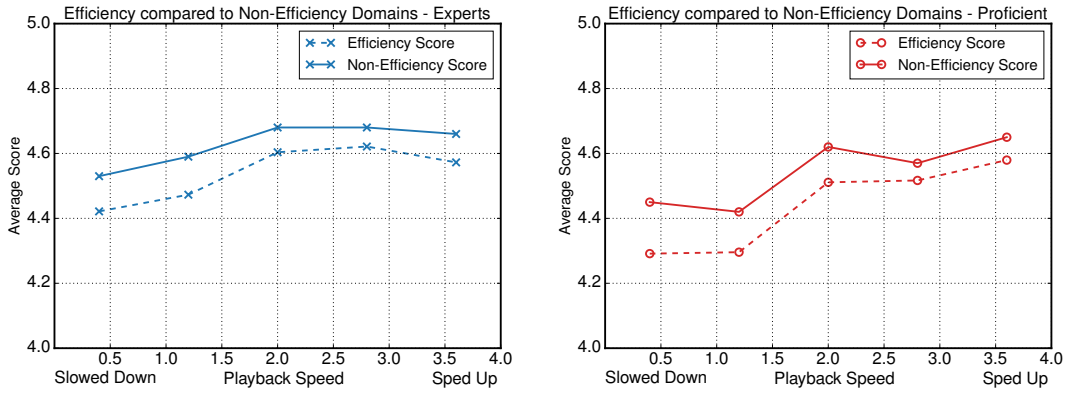


Figure 5.12: The efficiency subdomain as compared to the mean of the other GEARS subdomains, for expert and proficient surgeons.

5.2.5.1 Technical Skill Perception

The mean of the GEARS score for each group and video are shown in Fig. 5.11. For expert videos, each increase in the playback speed by 0.8x was associated with, on average, a 0.16-point increase in the GEARS score (95% CI: 0.10-0.22 point increase; $p < 0.05$). On average these scores appear to increase within a sublevel of the playback speeds around 0.4x to 2.0x, and then level out at all remaining playback speeds. For proficient surgeon videos, each increase in the playback speed by 0.8x was associated with, on average, a 0.27-point increase in the GEARS score (95% CI: 0.19-0.35 point increase; $p < 0.05$). Thus, while both experts and proficient surgeons experienced increased perceived technical skill as the playback speed was increased, the mean score obtained at the fastest video playback speed reached an almost equal skill level for both proficient and expert level performances. Fig. 5.12 shows the increase in the efficiency subdomain as well as the mean of the other four domains to visualize whether efficiency (seen as the most related to speed) is the only increasing domain. As shown, no major difference is apparent between the two types of domains.

5.2.5.2 Sub-Task Level Labeling

Figure 5.13 illustrates the difference in GEARS scores by non-expert crowd workers, when viewing the first minute of the video compared to the entire video. As shown, there is a noticeable difference in the scores given at these two levels. Most surgeons, when viewed for the entire 15 minutes, are rated slightly higher than when only the first minute of surgical activity is performed. Viewing just the label given to the performance (proficient or expert), by analyzing whether the video was above or below the median score in the group of 56 videos, a total of 11 previously labeled proficient level surgeons, and 11 previously labeled expert surgeons switched the label they were originally given, when only the first minute of surgical activity was evaluated.

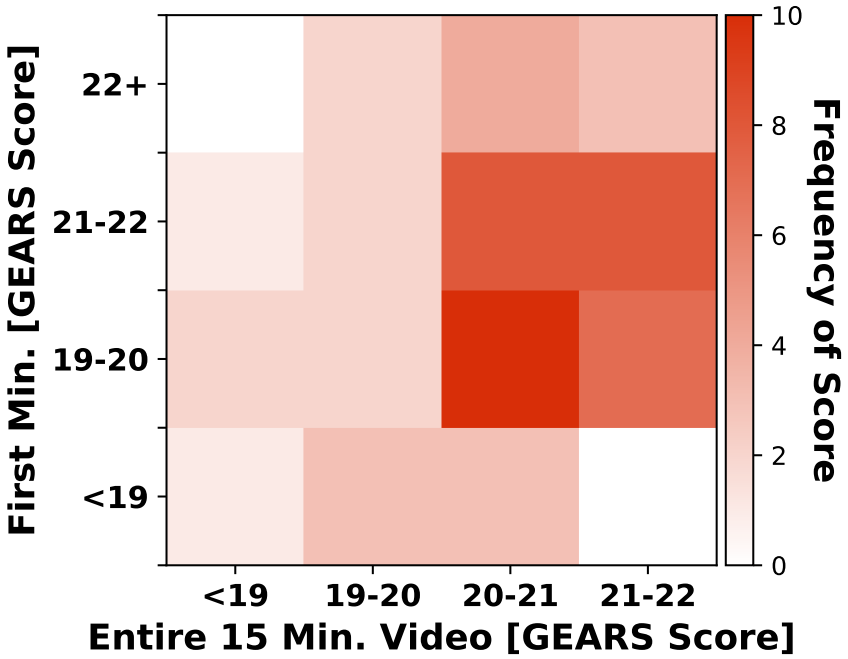


Figure 5.13: GEARS scores given to the entire 15 minute video of a performance, compared to only the first minute of the same video.

5.2.6 Conclusion

The results from the technical skill perception study give support to our initial hypothesis that increasing the video playback speed would increase the ratings of experienced surgeons. Now this evidence extends to real surgeries using robotic surgical performances. Surprisingly, however, we discovered that for sub-groups of expert level surgeons (proficient and expert) the increase in score happens at a quicker rate for proficient surgeons than for expert level surgeons. Increasing the playback speed of a slightly less than expert-level surgeon tends to "wash out" the minor mistakes they have made in the performance, effectively making the two groups appear more equally skilled at higher playback speeds. Additionally from analyzing the 'Efficiency' subdomain from the GEARS assessment, it appears that - surprisingly - crowds are also biased to give higher ratings in domains which are not associated with speed.

The results from the sub-task level skill labeling experiment show us that it may be necessary to have video evaluated in smaller duration segments, due to the notable disagreement with the skill level assigned to the entire video vs. just using the first minute. This lends support to the notion that a surgeon's technical skill may fluctuate in time (on the order of minutes) throughout a surgical procedure. This warrants further study of non-constant skill mapped to a single summative rating, as this would induce substantial label noise of computational skill evaluation using machine learning.

We conclude that increasing the video playback speed of performances of practicing surgeons in typical robotic surgeries results in increased scores as reported by non-expert crowds. This effect is surprisingly uniform across GEARS subdomains, even those which should be unaffected by speed. We further conclude that more studies should be done to investigate variance in time of sub-task level videos of surgical procedures, as the technical skill of a surgeon may fluctuate on the order of minutes or less during the

procedure.

Chapter 6

Robotic Surgery Readiness Study

This chapter introduces the RSR study and dataset, detailing the motivation, methods, and results from this study, which contains a reproduction of a journal article to be submitted to JAMA Surgery entitled "Robot-Assisted Surgery Readiness: Virtual Reality Warm-Up Prior to Robot-Assisted Surgery". In addition this chapter includes a reproduction of a journal article entitled "Temporal Variability in Surgical Technical Skill Evaluation" which used a subset of the RSR data and was submitted to the International Journal of Computer Assisted Radiology and Surgery in April 2020.

6.1 Robot-Assisted Surgery Readiness: Virtual Reality Warm-Up Prior to Robot-Assisted Surgery [93]

6.1.1 Abstract

Background: Surgical errors and complications have been linked to the volume of cases a surgeon experiences, and research has identified a distinct decrement of skill over intervals of inactivity. High stakes professions including athletics, aviation, and music have incorporated brief pre-procedural warm-ups to counter skills decay and elevate

readiness. The goals of this research were to identify the optimal virtual reality (VR) warm-up module to prime a surgeon for robotic surgery, and validate it in the operating room (OR) using video review and objective performance metrics in a randomized controlled setting.

Study Design: Surgical trainees and faculty at three medical centers were recruited for participation to complete VR training on proficiency tasks. Each surgeon was randomized to be in one of five groups. Members of each group performed six trial sessions of warm-up modules assigned to their group to prime them for surgery. The warm-up curriculum was chosen from the group with the best task time (TT) and Global Evaluative Assessment of Robotic Skills (GEARS) scores, obtained by video review. Robotic surgeons were randomized to either receive or not receive the chosen warm-up module before beginning surgery. Video of the first 15 minutes of the surgeons performance was captured in the OR. The GEARS assessment tool was used to review these videos and primary metrics were calculated to assess the validity of the curriculum, by both expert raters and non-expert crowd workers. Both a linear mixed effects model with a random intercept for each surgeon and a nonparametric modified Friedman test were used to analyze the effect of warm-up.

Results: Forty-four surgeons participated in the study to find the optimal warm-up curriculum. Group 1, a lone Running Suture module, was on average 31.3 seconds faster, and had the highest GEARS scores with the least variation. The lone Running Suture module was thus chosen as the optimal RSR curriculum. Thirty-four surgeons completed 432 surgeries of which 343 videos with corresponding tooltip instrument position data were recorded, making this the largest dataset of its kind. No statistically significant difference between the performance of the control and intervention groups was found. Further, no significant differences between treatment arms were found in any of the six kinematic and event outcomes used. Kinematic and event outcomes were

not significantly associated with GEARS scores as quantified by repeated measures correlation.

Conclusion: It appears that in surgical settings, performing a simulated VR warm-up procedure before undergoing robotic surgery does not have a significant effect on the surgical technical skills of robotic surgeons. This dataset is the largest of its kind, and will be beneficial to future research efforts involving surgical technical skill.

6.1.2 Introduction

The Institute of Medicine reported that medical errors account for as many as 98,000 deaths each year, atop a higher number of patient complications [5]. An estimated one third of these errors are surgical and the average American can expect to have 7 surgeries in their lifetime [94–96]. Although surgical simulations role in off-line teaching is accepted [97], simulation may also have a role in identifying decrements in surgical performance and offer a system for mitigating skill decay for direct patient care [98–101]. Perez et al. expanded on this premise and challenged researchers to not only focus on psychomotor skills, but also on cognitive and perceptual skills. They concluded that skill may decay at different degrees depending on the type of skill, and that training to mitigate decay may require targeted curricula based on the specific decay signatures [99]. These methods were described for laparoscopic surgery skills, yet no intervention has been studied for robotic surgeons or reintegrated military personnel with robotic surgery practices.

Robotic surgery has experienced considerable adoption in hospitals across the United States with the creation of the da Vinci surgical robot by Intuitive Surgical (Sunnyvale, CA) [100] (Fig. 6.1). With this new technology comes the need to apply innovative methods to elevate surgeon performance.

While several other checklist-type preparation procedures exist in the OR, there is

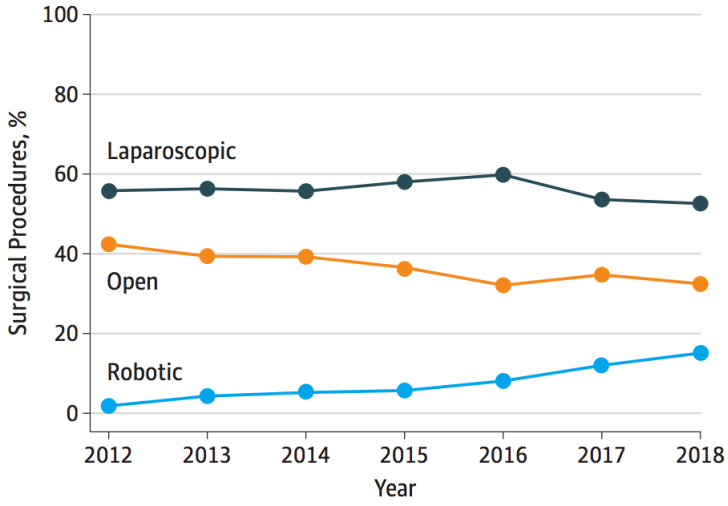


Figure 6.1: The popularity in robotic surgery has experienced a significant increase in recent years.

still no metric which measures how prepared a surgeon is following periods of inactivity [101]. It would be beneficial to quantify a measure of readiness to ensure a surgeon will perform at the optimal level of proficiency. It has been shown in other domains that after participating in a warm-up, task performance can be elevated [102]. Previous work additionally shows that a warm-up decrement (WUD) exists, in which the effects of warm-up begin decaying approximately 30 minutes after performing a warm-up task. [102–105]. These hypotheses have already found four factors which most affect WUD: (1) intervals between practice, (2) complexity and type of task, (3) strategies employed to maintain skills, and (4) individual differences [106, 107]. Using VR proficiency-based warm-up modules have been shown to most effectively combat skill decay [108–112].

The da Vinci Skills Simulator (dVSS), created by Intuitive Surgical (Sunnyvale, CA), replaces the physical robotic arms of a da Vinci robot with virtual representations of the tools to simulate a surgical environment (Fig. 6.2) [113]. The simulation software was created by Mimic Technologies Inc. in 2007, and allows surgeons for the first

Figure 6.2: The da Vinci Skill Simulator (dVSS) used for surgical simulation modules, created by Mimic Technologies.

time to practice surgery without any of the danger involved with real surgery (Mimic Technologies Inc). There are over 30 exercises available for practice with each of the da Vinci S, da Vinci Si, and da Vinci Xi. These tasks include suturing a wound and a typical ring-and-rail task for practicing dexterity.

Measuring surgical technical skill in an objective manner is exceedingly difficult, with the current gold standard being video review by expert surgeons. Previous work has shown that video review of surgeons by non-expert crowds is surprisingly accurate. When compared to expert surgeon review, crowds of non-experts were able to predict the pass/fail rate of surgeons with 100% accuracy in retrospective studies [1, 19]. In addition to these human perception-based approaches, computational and event metrics have also predicted surgical technical skill with varying degrees of accuracy. Metrics such as Task Time (TT), spectral arc length (SAL), camera use per minute (CUP), normalized angular displacement (NAD), rate of orientation change (ROC), and mean velocity (MV), have all been found to associate with surgical skill [13, 27, 39, 40, 114].

In this study we sought to find the optimal VR warm-up module to properly prime surgeons for robotic surgery, and validate the module in the operating room (OR) by measuring objective performance metrics and obtaining video review of participants who were randomized to either participate in the module or not prior to surgery.

6.1.3 Methods

6.1.3.1 Aim 1

6.1.3.1.1 Subject Recruitment Surgeons and trainees were recruited after Institutional Review Board (IRB) approval (UW IRB # 41730), from urology, gynecology and general surgery at the University of Washington Medical Center (UWMC), Madigan Army Medical Center (MAMC), and the Florida Hospital/Nicholson Simulation Center. Each of the subjects were asked to complete a standard demographics questionnaire to learn their experience and any factors impacting responses to warm-up simulations. The information obtained included:

- Unique Subject Code (e.g. UWMC03)
- Age
- Gender
- Years in Training
- Years since graduated training
- Surgical specialty
- Robotic surgery exp. (# cases as primary surgeon)
- Laparoscopic surgery exp. (# cases as primary surgeon)
- Handedness

- Musical experience
- Date of return since last deployment
- Duration of deployment
- Surgical role while deployed (if applicable)
- Time since last performed robotic surgery or procedure

6.1.3.1.2 dVSS Modules The modules used for surgeon priming needed to be complex and effective enough at stimulating the surgeon in a way that most closely corresponds to movements present in a surgical environment. The criteria used to select modules were: (1) prior validation in the literature, and/or 2) involvement of multiple skills simultaneously, and/or 3) containing content that closely simulates actual surgery. Following these criteria, we chose to use the Ring and Rail 2, and Match Board 3 modules from the EndoWrist Manipulation 2 category. Ring and Rail 2 has previously been found to effectively discriminate against surgical skills, and Match Board 3 has been found to be the most difficult of the dVSS modules among experienced surgeons [109, 115]. From the Needle Driving category, Suture Sponge 3 was chosen as it has been found to vary greatly between experience, intermediate, and novice performers [115]. Running Suture was the final module chosen, and the criterion module, as it most closely represents the actions in a real surgical suturing procedure. Each of these four modules can be completed in approximately 2-3 minutes, making them short enough to be feasibly used as a component of a final curriculum (Fig. 6.3).

6.1.3.1.3 Proficiency Testing All recruited participants did not necessarily have the same levels of experience or skill. To train a benchmark of technical skill, each surgeon was required to complete the Intuitive Surgical da Vinci didactic web-based curriculum. Previous research has used this curriculum as a method to familiarize

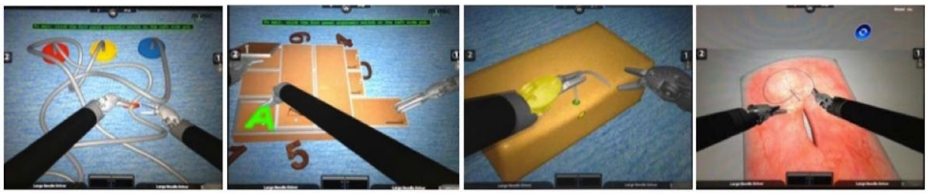


Figure 6.3: The four dVSS warm-up modules used for aim 1 of the RSR study including Ring and Rail 2, Match Board 3, Suture Sponge 3, and Running Suture (left to right).

participants with the console and train surgeons to adequate proficiency levels [111]. Two expert robotic surgeons completed the four modules used, until they made no errors. These performances were used as the benchmark for expert skill.

The four dVSS modules were split into five groups with different combinations of modules, to test which module or module series was most efficient for priming surgeons (Fig. 6.4). The modules in each of these five groups are shown in Table 6.1. Participants were randomized to be in one of the 5 groups, and within each group, half were randomized to perform the tasks monthly and half performed bi-weekly. After 8 weeks of performing at one frequency, the participants were then swapped to the opposing frequency group, to allow for certain baselines in statistical models to control for intervals of inactivity. For groups with more than 1 module in a series, a 30 second period was inserted between each module performed, making the total simulation time approximately 3-12 minutes. After performing the module or module series, each participant performed the criterion module (Running Suture) in which they were assessed based on their performance by the two expert surgeons.

6.1.3.1.4 Video Review Two expert surgeons reviewed videos of the performances and rated the criterion suturing procedures. The GEARS assessment tool, being the most popular of the assessment tools for robotic surgery, was used as the evaluation

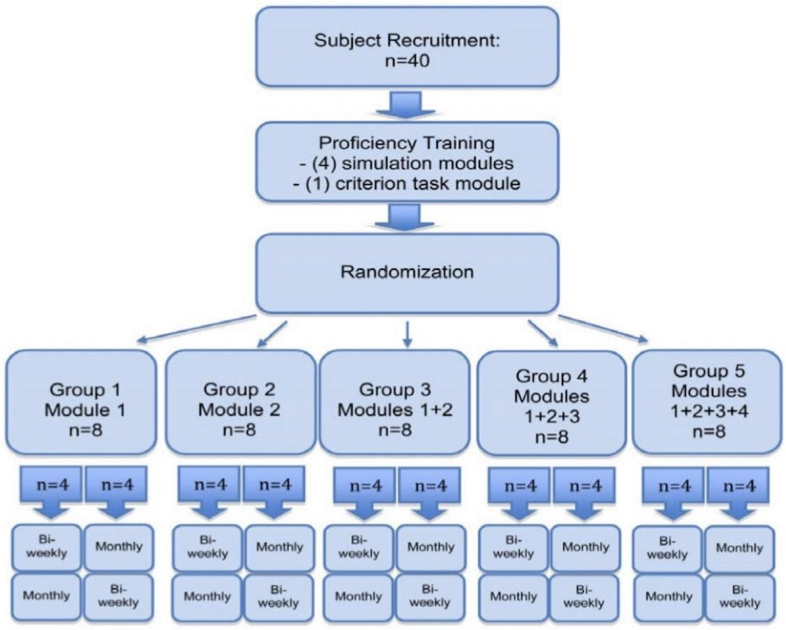


Figure 6.4: The randomization setup for the 5 warm-up groups in aim 1 of the RSR study.

Simulation Group	Group 1	Group 2	Group 3	Group 4	Group 5
Running Suture	X		X	X	X
Ring and Rail		X	X	X	X
Match Board				X	X
Suture Sponge					X

Table 6.1: The five groups of warm-up modules used for aim 1. Each group consisted of different combinations of warm-up modules shown to help elevate skill.

method [89]. The GEARS evaluation consists of 6 Likert-scale domains which are relevant to robotic surgical technical skill. These domains are: depth perception, bimanual dexterity, force sensitivity, autonomy, efficiency, and robotic control. Each of these domains are given a score between 1 and 5, then added for a cumulative score between 6 and 30, with a higher score corresponding to a more skilled surgeon. Autonomy was removed from consideration for this study, as it can be used with a lone video review, leaving a total possible score of 5-25.

6.1.3.1.5 Statistical Models The primary goals of Aim 1 were to evaluate (1) the magnitude of skills decay between the two standardized assessment intervals, and (2) the curriculum which optimizes the participants GEARS scores and results in the smallest deviation from performer baseline performance.

The general model used was a linear mixed effects model which accounts for interval assessment period, timing of assessment, as well as the contrast between the different curriculum groups. This model included a random intercept to account for clustering of study participants, and a fixed effect to adjust for assessment time between different trials. To account for the average percentage of skill decay between groups, the contrast in GEARS scores between groups performing monthly and groups performing bi-weekly was accounted for in the model. The module which yielded the highest average combined GEARS score after accounting for all of these factors was used as the optimal warm-up curriculum for Aim 2.

6.1.3.2 Aim 2

6.1.3.2.1 Subject Recruitment Subjects from Aim 1 were allowed to participate in Aim 2 of the study, in addition to newly recruited participants which were approved by the IRB. Each newly enrolled subject in the second Aim completed the demographics

questionnaire discussed in the section for Aim 1.

6.1.3.2.2 Randomization and Warm-Up Protocol Individual cases from each recruited surgeon were randomized to either receive the warm-up curriculum or not receive it, with each surgeon acting as their own control case. These assignments were stored centrally at the University of Washington Center for Biomedical Statistics (CBS) and distributed using a REDCap-based (http://project-redcap.org/) web delivery system. Each sites research coordinators ensured adherence to the warm-up protocol. Surgeons who were randomized to receive the warm-up were brought to a dVSS tower to complete the VR module selected in Aim 1. Participating surgeons completed between 1-40 procedures as a part of this study. Each sites project coordinator worked with operating room staff to familiarize the operative team with the projects protocol to minimize disruptions to the daily operation routines.

6.1.3.2.3 OR Data Capture Endoscopic video and robotic instrument position was captured for each procedure using the Intuitive Surgical dvLogger system, which captures data directly from the robots API. Capturing this data allowed for using the criterion video with the GEARS assessment tool by expert surgeons and non-expert crowd workers, afterwards. This also allowed performance metrics to be computed using the obtained kinematic data from the criterion surgery.

6.1.3.2.4 Data Aggregation and Manipulation All video and tooltip data was sent in external hard drives to UMN for analysis. Files were then parsed into 'SI' or 'XI' labels based on which da Vinci platform was used.In several on-screen modules used to display which surgeon had control of which arms of the robot, the video in SI cases displayed personally identifiable information. These displayed modules were censored using OpenCV and Python to create black boxes over these areas.

6.1.3.2.5 Video Segmentation The videos in each surgical case were manually searched to find the timestamps at which the surgeon started operating. An effort was made for each video to obtain the first 15 minutes of surgery from the criterion surgeon, as this was the timeframe needed to see the largest effect from the warm-up module. Throughout several cases, the criterion surgeon was switching back and forth with a partner, requiring editing and splicing together several sections of video together to arrive at a 15 minute timeframe. An additional effort was made to remove all sections of video in which the endoscope leaves the patient's body, capturing possibly identifiable information. This video was then uploaded as "unlisted" to Youtube, to allow sharing videos with faculty surgeon reviewers.

6.1.3.2.6 Objective Metric Calculation The kinematic tooltip data was parsed using Python [87] and Intuitives internal API for decrypting the data. The dvLogger split surgeries chronologically into several directories in which a surgery could comprise over 100 separate directories, depending on the length of the surgery. A script was written to store each bin's kinematic data individually in a compressed format for later analysis. An additional script was written to find the time intervals in the surgery when the arm's were moving to calculate tool path length. The information obtained from the raw data contained X, Y, Z coordinates in meters, along with a UNIX timestamp in milliseconds. The kinematic data and video were usually not well-synced so an effort was made to align the criterion portion of the video with the kinematic tooltip data by calculating the percentage of the video completed with the percentage of timestamps completed in the kinematic data, to get the closest possible match between the two modalities. These data were used to obtain summary metrics based on periods when the surgeon was performing, including rate of orientation change (ROC), spectral arc length (SAL), normalized angular displacement (NAD), mean velocity (MV), laterality

(LR), and camera use per minute (CUP).

6.1.3.2.7 Video Review For faculty surgeon review, a web-based platform was created to distribute the processed endoscopic videos accompanied by the modified GEARS questionnaire to faculty surgeons and non-expert crowds. A subset of the total videos were used for faculty surgeon review, attempting to obtain reviews for as many videos as possible. Each GEARS score was recorded for each performance to analyze the intraclass correlation (ICC) among ratings for the same video. Statistical calculations were computed using R [92]. C-SATS Inc. was additionally used to compile quick and reliable reviews of technical skill for each video, obtained from crowds of non-experts, to offset the limited availability of faculty surgeons.

6.1.4 Results

6.1.4.1 Aim 1

Forty-one participants completed the study in Aim 1, although the original objectives were to only enroll forty, with demographics from these surgeons shown in Fig. 6.5. As can be seen in Fig. 6.6, the average TT for the different groups had a downward trend in which the surgeons became quicker as they proceeded through sessions. Clearly, the Ring and Rail 2" lone module, group 2, was the least effective at lowering TT while also, as seen in Fig. 6.7, resulted in the lowest GEARS scores. Groups 1 and 5 resulted in the most favorable TT and GEARS scores. Overall group 1 participants, the participants only priming with the "Running Suture" module, had a higher GEARS score (2.1 points higher +/- 0.86, $p = 0.02$). Compared to group 2, group 1 was on average 31.3 seconds faster (+/- 14.02 seconds, $p = 0.03$). GEARS scores from groups 1 and 5 were not significantly different from each other ($p = 0.47$ and $p = 0.70$), although group 1 had the least variation in GEARS scores and TT of all groups. The criterion task, "Running

	Florida	Madigan	Washington	Total
Target Enrollment	20	20	30	70
Total Randomized	**5 (25)**	**29 (145)**	**12 (40)**	**46 (66)**
Completed Participation	**3 (15)**	**27 (135)**	**11 (37)**	**41 (59)**
Consented*	7 (35)	38 (190)	29 (97)	74 (106)
Gender				
Female	1 (14)	10 (26)	9 (31)	20 (27)
Male	6 (86)	28 (74)	17 (59)	51 (69)
Other	0 (0)	0 (0)	0 (0)	0 (0)
Seniority				
Attending	7 (100)	12 (32)	3 (10)	22 (30)
Senior	0 (0)	9 (24)	9 (31)	18 (24)
Junior	0 (0)	17 (45)	14 (48)	31 (42)
Specialty				
Urology	3 (43)	12 (32)	12 (41)	27 (36)
Gynocology	1 (14)	9 (24)	2 (7)	12 (16)
General	2 (29)	8 (21)	10 (34)	20 (27)
Thoracic	0 (0)	0 (0)	0 (0)	0 (0)
Cardiothoracic	0 (0)	0 (0)	0 (0)	0 (0)
ENT	1 (14)	9 (24)	2 (7)	12 (16)
Handedness				
Left	1 (14)	1 (3)	1 (3)	3 (4)
Right	5 (71)	35 (92)	23 (79)	63 (85)
Ambidextrous	1 (14)	2 (5)	2 (7)	5 (7)
Musical Experience	2 (29)	22 (58)	19 (66)	43 (58)
Deployed	0 (0)	7 (18)	0 (0)	7 (9)

Figure 6.5: Enrollment demographics for participants in aim 1 of the RSR study.

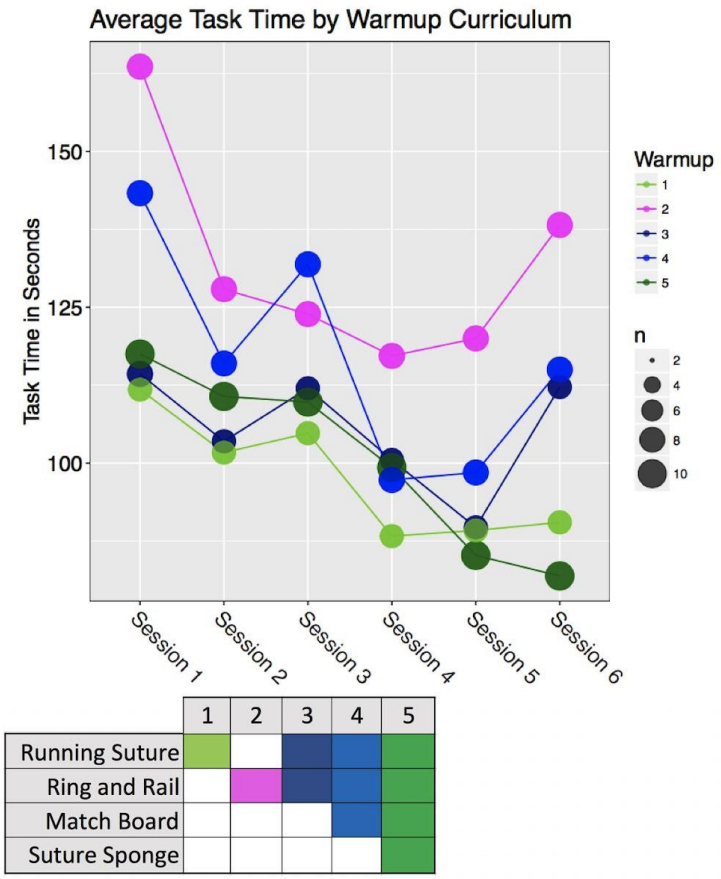


Figure 6.6: The task time (TT) for the five different warm-up module groups, throughout the six completed sessions.

Suture", thus was chosen as the warm-up curriculum to use before surgical procedures.

6.1.4.2 Aim 2

Forty surgeons were enrolled in Aim 2, from UWMC and MAMC, with 34 surgeons having completed at least one surgery. Of the 34 surgeons, 30 answered the demographics questionnaire, with results shown in the Fig. 6.8. Video of 432 RAS were performed and recorded. Due to various technical malfunctions with the recording technology, the kinematic data for 89 cases were not recorded, leaving 343 videos having corresponding

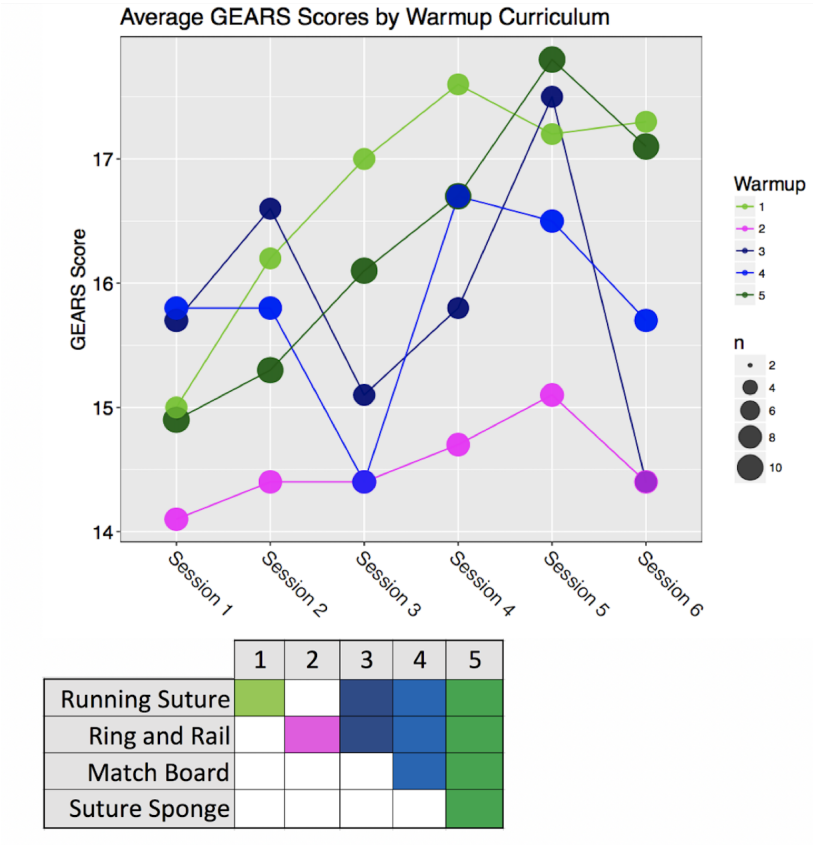


Figure 6.7: The GEARS scores for the five different warm-up module groups, throughout the six completed sessions.

	Madigan	Washington	Total
Sessions with Video/Kinematic	17	326	343
Completed Sessions	41	391	432
Consented*	12 (200)	22 (110)	34 (131)
Deployed Currently*	0 (0)	0 (0)	0 (0)
Lost to Follow-Up*	0 (0)	0 (0)	0 (0)
Gender			
Female	1 (8)	8 (36)	9 (26)
Male	10 (83)	11 (50)	21 (62)
Other	0 (0)	0 (0)	0 (0)
Specialty			
Urology	2 (17)	8 (36)	10 (29)
Gynocology	0 (0)	9 (41)	9 (26)
General	6 (50)	1 (5)	7 (21)
Thoracic	0 (0)	0 (0)	0 (0)
Cardiothoracic	1 (8)	0 (0)	1 (3)
Handedness			
Left	0 (0)	0 (0)	0 (0)
Right	11 (92)	19 (86)	30 (88)
Ambidextrous	0 (0)	0 (0)	0 (0)
Musical Experience	3 (25)	6 (27)	9 (26)
Deployed Previously	9 (75)	0 (0)	9 (26)

Figure 6.8: Enrollment demographics for participants in aim 2 of the RSR study.

tooltip position kinematic data.

6.1.4.2.1 GEARS Review Outcome Ratings by seven faculty surgeons were obtained for 45 videos, with two raters rating all videos and the other five rating from 7 to 44 videos each. ICC for expert ratings was 0.3825 (95% CI from 0.2473 to 0.5366), indicating poor agreement. When the two raters with the greatest numbers of missing observations were removed, ICC was 0.4502 (95% CI from 0.3026 to 0.6043), still indicating poor agreement. ICC estimates and their confidence intervals were based on an individual ratings, consistency, 2-way random-effects model for incomplete datasets

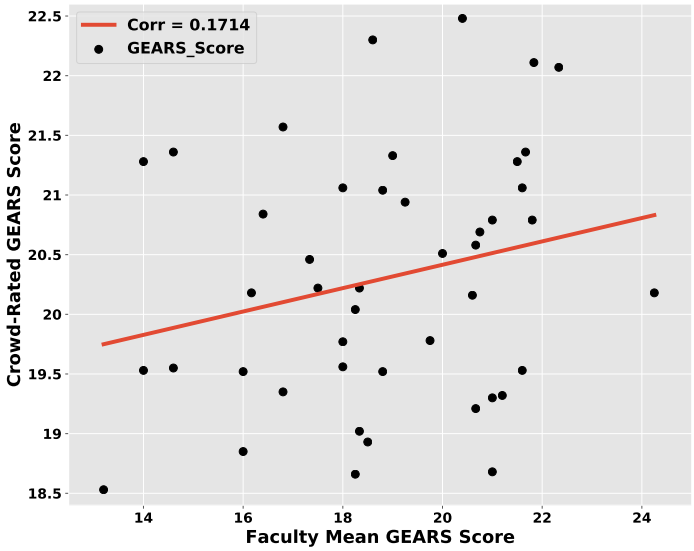


Figure 6.9: The correlation of the GEARS scores from faculty surgeons and C-SATS scores was weak, with a value of 0.1714, 95% CI from -0.1287 to 0.4426.

(using R package irrNA, Brueckl and Heuer, 2018). Crowdsourced GEARS scores supplied by C-SATS were moderately (but not significantly) correlated with the mean of the expert ratings for the same case ($r(43) = 0.1714$, 95% CI from -0.1287 to 0.4426, Fig. 6.9).

We assessed the effect of the intervention on GEARS score, our primary outcome, using both a linear mixed model with a random intercept for each surgeon ($p = 0.9819$) and a nonparametric modified Friedman test ($p = 0.3441$), with both finding no significant difference between cases receiving or not receiving the intervention. Robot type (XI vs SI) was also not significantly related to GEARS score ($p = 0.2365$), nor was any robot by treatment interaction detected ($p = 0.6694$, Fig. 6.10).

6.1.4.2.2 Objective Metric Outcomes We also examined the effect of the intervention on six kinematic and event outcomes (mean velocity (MV), rate of orientation change (ROC), normalized angular displacement (NAD), spectral arc length (SAL),

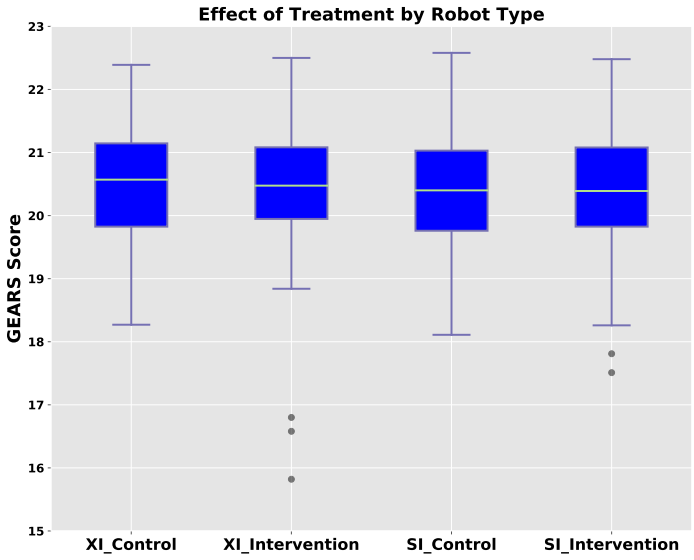


Figure 6.10: The effect of Si and Xi da Vinci platforms for surgeons who performed and did not perform a warm-up module before surgery. No significant difference exists between these four groups.

camera use per minute (CUP), and laterality (LR)) using both linear mixed models and a nonparametric modified Friedman test, finding no significant differences between treatment arms. We used the Holm-Sidak procedure to maintain familywise alpha at .05.

Kinematic and event outcomes were not significantly associated with GEARS score as quantified by repeated measures correlation (using R package rmcorr [116]), (Table 6.2, and 6.3.

The median time between surgeries for surgeons in our study was 20 days. We did not find evidence for a difference in GEARS score ($p = 0.8331$) or treatment effect ($p = 0.5997$) associated with greater time lapse between surgeries (Table 6.4.

A linear mixed model including both a random intercept and a random slope for the effect of the treatment on each surgeon was used to quantify the range of effects the intervention had on different surgeons. Treatment effects for individual surgeons on

Metric	Estimated β	P-Value	P-Value Threshold
MV	0.0587	0.397	0.009
NAD	-0.063	0.426	0.010
SAL	0.059	0.655	0.013
CUP	0.027	0.766	0.017
LR	-0.009	0.790	0.025
ROC	0.021	0.888	0.050

Table 6.2: Results from linear mixed model tested with kinematic and event outcomes.

Metric	P-Value	P-Value Threshold
MV	0.3536	0.017
NAD	0.2909	0.010
SAL	0.3132	0.0127
CUP	0.6797	0.050
LR	0.4898	0.0253
ROC	0.1595	0.0085

Table 6.3: Results from Friedman test with kinematic and event outcomes.

Metric	Correlation	95% Upper CI	95% Lower CI	P-Value
MV	0.078	-0.031	0.187	0.161
NAD	-0.056	-0.165	0.054	0.315
SAL	-0.022	-0.132	0.088	0.692
CUP	0.079	-0.031	0.188	0.158
LR	-0.010	-0.119	0.100	0.864
ROC	-0.013	-0.123	0.097	0.811

Table 6.4: Correlation values for each of the metrics measured with GEARS scores.

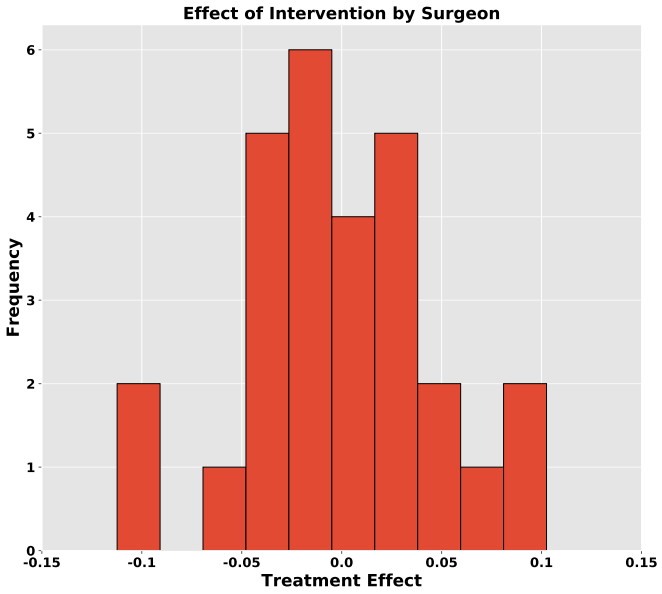


Figure 6.11: The effect of performing a warm-up module for each each surgeon in the RSR study. With a clear normal distribution, no significant effect was measured.

GEARS score ranged from -.102 to .094 points, which we judged not to be clinically meaningful, thus precluding the possibility of identifying subgroups of surgeons who benefited from the intervention to any practically significant degree (Fig. 6.11).

6.1.5 Discussion

No statistically significant difference in perceived technical skill was observed between robotic surgeons who perform a warm-up suturing module and those who do not. Tests controlling for individual surgeons, Si vs. Xi da Vinci robot platforms, as well as time since last surgical performance, all found no evidence that performing the RSR curriculum prior to surgery elevated surgical readiness, as measured through both kinematic event metrics and faculty video review.

Efforts from this study led to the creation of the largest dataset (to the authors knowledge) of RAS with corresponding kinematic tooltip data to date. This could lead

to breakthroughs in artificial intelligence models which map surgical technical skill to robotic surgery performance and video features. Large datasets are necessary for proper training of machine learning models which can both quantify technical skill and perform semantic segmentation. The rich, multimodal detail provided by this dataset, shown through both high definition video and the abundance of kinematic tooltip data should prove invaluable for future research.

6.1.6 Conclusions

Performing a VR warm-up module prior to robotic surgery does not significantly improve the technical skill or readiness of a robotic surgeon. The dataset created in this study will prove invaluable for future research including surgical skill classification, semantic segmentation, automated performance metric calculation, and surgical simulation research.

6.2 Temporal Variability in Surgical Technical Skill Evaluation [117]

6.2.1 Abstract

Purpose: Summary score metrics, either from crowds of non-experts, faculty surgeons, or automated performance metrics (APMs), have been trusted as the prevailing method of reporting surgeon technical skill. The aim of this paper is to learn whether there exist significant fluctuations in the technical skill assessments of a surgeon throughout long durations of surgical footage.

Methods: A set of 12 videos of common robotic surgery cases from human patient robotic surgeries were used to evaluate the perceived technical skill at each individual minute of the surgical videos, which were originally 12-15 minutes in length. A linear mixed effects model for each video was used to compare the ratings of each minute to the overall mean for the entire video in order to learn whether a change in scores over time is more significant than typical intrarater variation.

Results: Modeling the change over time of the Global Evaluative Assessment of Robotic Skills (GEARS) scores significantly contributed to the prediction model. This gives evidence that significant changes in the technical skill over time are noticeable in the model.

Conclusion: The findings from this research raise questions about the optimal duration of footage needed to be evaluated to arrive at an accurate rating of surgical technical skill for longer procedures. This may imply non-negligible label noise for supervised machine learning approaches. In the future, it may be necessary to report a surgeon's skill variability in addition to their average score to have proper knowledge of a surgeon's overall skill level.

6.2.2 Introduction

Methods for assessing the technical skills of surgeons is paramount to ensuring to the public that surgeons are safe and effective. For years summary scores or metrics have been used as the main method to report surgical skill. The most popular of these has been to use a likert-scale scoring metric, from either non-expert crowds or faculty surgeons. Past research has shown that crowds of non-experts concord with surgeon raters in evaluating technical skill [19]. Automated performance metrics have also been used, in which surgical events or streaming kinematic data have been used for computation of various metrics, suggesting superior objectivity [41]. As it is known that surgical skill is related to patient outcomes, and medical errors are the third leading cause of death in the United States, for which surgical errors contribute a large part, it remains important to accurately assess and report surgeon skill [4], [72]. Much progress has been made by using statistical and machine learning models in the past with the goal of classifying surgeons into skill levels of 'novice' and 'expert' [73], but it remains a difficult computational task. For the largest statistical power this usually leaves the most reliable approach being to obtain either surgeon or crowd evaluations from video [22].

One of the most popular robotic surgical skill assessment metrics is the Global Evaluative Assessment of Robotic Skills (GEARS), which is the most common objective assessment tool for robotic surgery [89]. The subdomains in this metric include: bimanual dexterity, efficiency, depth perception, force sensitivity and robotic control. Scores of 1-5 for each of these subdomains are totaled, for a cumulative score of 5-25, with higher scores belonging to more skilled surgeons. There is an additional domain in GEARS named Autonomy which is not used here. Autonomy is typically used when evaluating a surgeon's ability to work independently, which can't be accurately assessed through video alone.

At least two confounding factors arise: (1) limitation of human attention span of the raters which inhibits reliability of their ratings and (2) the fluctuations of "true technical skill", naturally exhibited by a surgeon through time, when videos of a longer duration are used. It is unclear whether ratings remain reliable as durations scale up. The former is akin to *measurement noise* and is an attribute of the method to measuring skill - be it human ratings or computed computationally. The latter is akin to *process noise* and is an attribute of the surgeon's activities and tool-tissue environment interactions. Studies of human attention span have found that people viewing video lectures on average experienced significant decreases in attention and even stopped viewing after only about 6 minutes. This study additionally found that including interactivity elements to video lectures led to an increase in watch time by at least 20% [118]. Given that evaluating surgeons requires an even higher degree of focus, using videos which are longer than 5 minutes may be detrimental to receiving accurate evaluations. Additionally, when making judgments, crowds can be sucsceptible to a contrast effect bias in which wide ranges in performance can lead to wildly different evaluations [119]. This could mean a performance of a novice who performed unusually well in the last segment of a video may receive unreliably high ratings. The natural fluctuations in technical skill remain largely unexplored, though hinted at in [91], [81].

We hypothesize that statistically significant degrees of temporal fluctuations of perceived skill exist between smaller segments of surgical performances, which differ from previously obtained summary scores of the longer duration videos.

6.2.3 Methods

6.2.3.1 Dataset

This study utilized 12 videos from the Robotic Surgery Readiness (RSR) study. The goal of that study was to determine whether pre-operative warm-up on virtual reality tasks had a measurable improvement in surgical technical skills among practicing surgeons (no novice trainees) after typical periods of surgical inactivity, in a controlled-randomized trial. This dataset contains 343 videos of surgical procedures consisting of live robotic surgeries, which were performed by forty attending surgeons and trainees in urology, gynecology, and general surgery at the University of Washington Medical Center and the Madigan Army Medical Center. These robotic surgeries were performed using da Vinci surgical robots, created by Intuitive Surgical (Sunnyvale, CA). Each video was manually edited to include roughly the first 15 minutes of surgical activity performed by the criterion surgeon. This beginning portion of the surgery contained similar actions among the different surgery specialties, such as suturing and cutting actions. A GEARS score for each 12-15 minute performance was also previously obtained from crowd evaluation.

The range of scores in all RSR videos was fairly small, with most lying between 20-22 out of 25 (only 15% of the full range possible). This was the case as all surgeons assessed were practicing faculty and no trainees were involved in these videos. To obtain the largest possible range of skill from this dataset, 6 performances from the top quintile of scored performances and 6 from the bottom quintile of performances were used and given labels of 'expert' and 'proficient', respectively, keeping in mind that all surgeons in this dataset are highly-skilled, or they wouldn't be allowed to operate in live robotic surgery procedures [70]. Additionally, half of each group of videos were performances in which the surgeon participated in a simulated surgery warm up procedure beforehand,

Group	Sub-Group	Videos (N)	Prostatectomy	Hysterectomy	Other
Expert	Warmup	3	1	1	1
	Control	3	1	0	2
Proficient	Warmup	3	2	0	1
	Control	3	1	1	1

Table 6.5: Summary information for 12 videos used from the RSR study.

whereas the other half started surgery without warming up. This stratified formatting is illustrated in Table 6.5.

6.2.3.2 Crowd Evaluation

Amazon Mechanical Turk was the crowd-sourcing platform used for this study, in which each non-expert crowd worker was paid an average of $0.50 to watch and evaluate each of a series of one minute videos. A web domain was created for which Turkers would be redirected, where they submitted a consent form and were asked GEARS questions about videos. Each 12-15 minute video performance was segmented into smaller videos having a duration of one minute, using FFMPEG, an open source video editing tool [85]. The order of these videos were randomized so that crowds were viewing a non-chronological ordering. They were asked to provide a GEARS score for each minute long video and proceeded to evaluate all remaining videos for the same performance.

A few extra elements were added to the user interface in an effort to increase interactivity and obtain better data quality. Once a crowd worker gave consent to participating in the study, they were shown a video with two performances, side by side, of a novice and expert performer of a laparoscopic surgical training task. The crowds were asked to evaluate which surgeon was better, and to what degree. After evaluating this task, the series of robotic surgery procedures began. After each GEARS likert-scale question

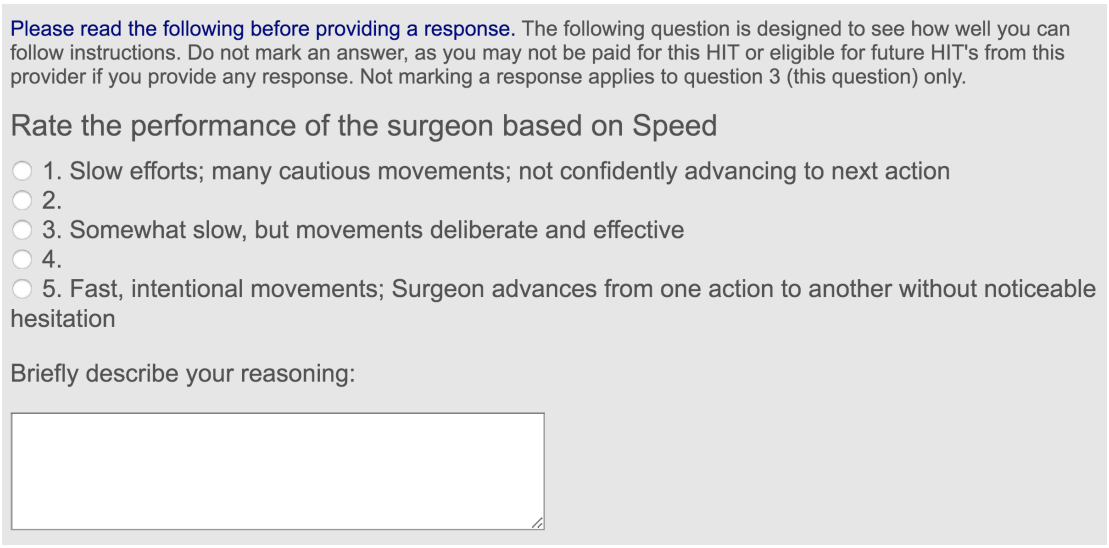


Figure 6.12: The attention question inserted to the GEARS questionnaire, for quality assurance purposes. Note that skimming the instructions likely results in incorrect answers.

prompt, a text box was inserted, asking the raters to provide reasoning for their decision in each subdomain.

Finally, an additional question, which is not part of the standard GEARS questionnaire was inserted, asking for a rating of the 'speed' of the surgeon, shown in Fig. 6.12. Text in the prompt stated that the worker should not answer this question. This 'attention question' served as a mechanism to query which turkers were focusing on the questionnaire vs. those quickly finishing to be paid. All ratings which included answers to this 'attention question' were removed from analysis.

6.2.3.3 Statistical Analysis

A linear mixed effects model for each video was used to analyze the outcome of GEARS scores assigned by each rater at each timepoint. A random intercept was estimated for each individual rater to model rater-by-rater variability in scores. Fixed

effects were also estimated for each minute of video in order to quantify differences in surgical skill over time. With each model, we tested the hypothesis that the mean GEARS score was the same at each timepoint to determine if there were significant changes in skill over time for each surgeon. All statistical work was accomplished using R 3.6.2 [92], with data manipulation and visualization computed in Python 3.6 [87].

6.2.4 Results

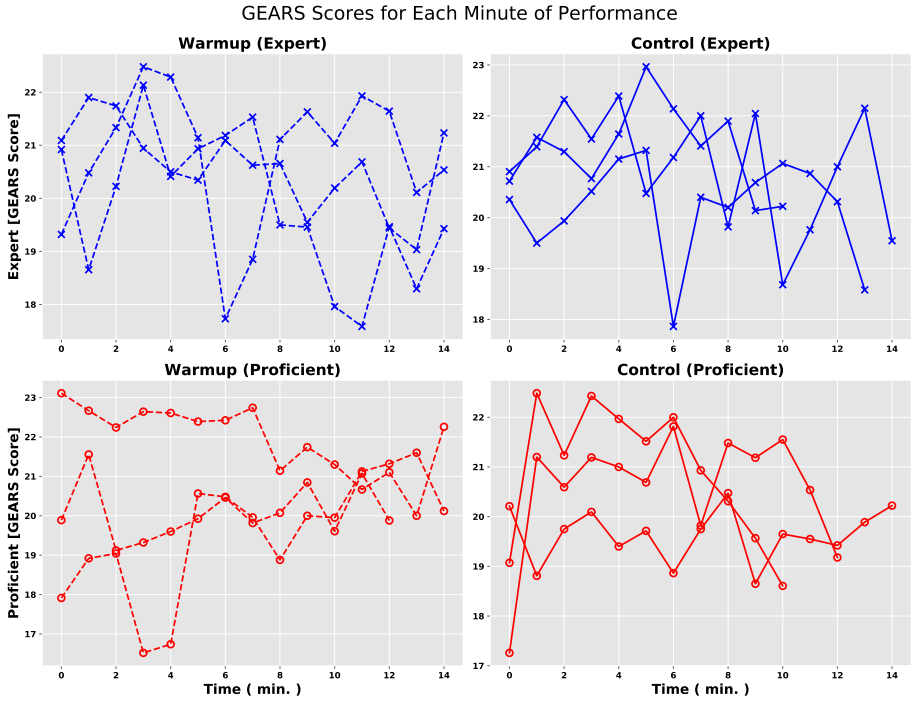


Figure 6.13: All mean crowd evaluations from each proficient (red) and expert (blue) surgeon at each minute of the performance. Surgeons previously randomized to the control (solid lines) group normally started surgery without any intervention or pre-operative warm-up used. Warm-up (dashed lines) group surgeons reviewed a virtual reality warm-up module prior to surgery. (The warm-up hypothesis from the original randomized study is not being tested or evaluated in this research, only temporal variation in ratings is.

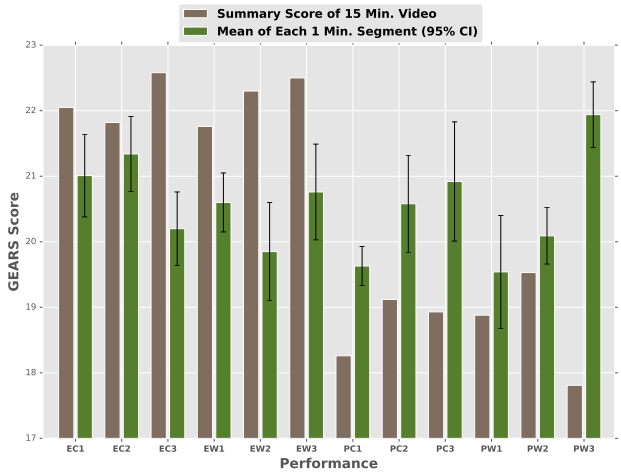


Figure 6.14: Comparison of previously-obtained scores for 15 minute video (95% CI) and average score for evaluation of 1 minute segment. (PW = Proficient/Warmup; PC = Proficient/Control; EW = Expert/Warmup; EC = Expert/Control)

The mean of the GEARS score for each expert and proficient surgeon are shown in Fig. 6.13. No identifiable relationship or trend in any of the 4 subplots is immediately apparent. However, significant differences in the average score over time were detected for ten of the twelve videos ($p = 0.001891$ to $2.531e-14$) using linear mixed effects models. For two videos, the effect of time was not significant (proficient control video 1, $p = 0.61794$, and proficient warm-up video 2, $p = 0.2618$).

A bar plot comparing the previously obtained summary score for each surgeon, grouped with the average of each 1 minute segment, is displayed in Fig. 6.14. Several of the previously obtained summary scores are outside the confidence interval band from the average of the one minute scores. Overall we saw each full duration video's 1-minute GEARS average scores vary by 1.02 ± 0.28 points.

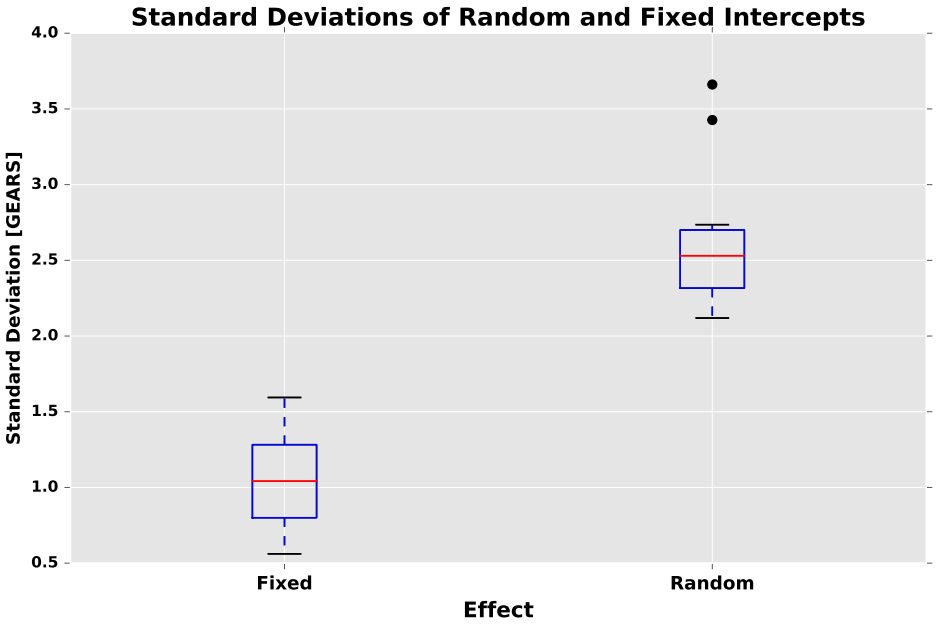


Figure 6.15: Standard deviation from the LME model for random effects from each individual crowd rater, compared with the fixed effect of time. The rater effect is clearly larger, although time is still significant in predicting outcomes.

6.2.5 Conclusion

The results from the video segmentation study support our initial hypothesis that perceived surgeon skill may fluctuate throughout surgical procedures. If we consider the limitations of human attention span and time-varying fluctuations in skill for a 15-minute video as contributing factors, it appears that human attention span limitations have a larger effect than natural variation of "true" technical skill in time (i.e. measurement noise may be 2-3x larger than process noise). We found no compelling evidence that the underlying human skill (process) is constant for a 15-minute duration in typical surgery. Perhaps crowds may be biased into giving scores that more closely reflect events they remember as being particularly good or bad. This research may give credit to alternative surgical skill reporting methods, as opposed to giving static scores

and labels. One alternative to this could be providing the mean of shorter segments of videos, as well as the amount of deviation in score throughout the performance, as a confidence interval. This would convey both the typical skill of a surgeon in addition to how consistent their skill is through time.

This study includes a few limitations. There was no semantic segmentation of these tasks, only temporal "chunking" into one minute segments. Although videos of real surgeries were used, the data used was fairly sparse, with only 12 videos of 15 minute surgeries. It is also not clear if the one minute duration used for segmentation is optimal. Shorter or longer segments of surgery may be more beneficial for accurate skill evaluation. In this work, the one minute ratings were obtained 3 months after the full duration videos were rated. This may raise concerns about reliability (and comparability of data) of crowd ratings from studies separated by large time intervals.

Using these alternative reporting methods could aid supervised machine learning models in the future by reducing potential label noise. Static scores may lead to noisily-labeled data sets, producing poorly trained classification models. Further study should consider optimal duration for human raters and a more rigorous analysis of the natural fluctuations of human technical skill.

Chapter 7

Conclusion and Discussion

This thesis presents work pertaining to multiple facets of how computational techniques in evaluating surgeon skill are affected by speed and task time. This work includes two research studies with conclusive evidence of biases to speed in skill evaluation efforts, as well as analysis surrounding how surgeon's hand movements correlate to expert skill and specific techniques which may be incorporated to classify surgical skill. In addition, this work includes findings from a new robotic surgery dataset in which it was found that performing a running suture virtual reality simulation prior to a robotic surgery did not lead to an elevated level of robotic technical skill. A secondary finding from this dataset shows the existence of temporal fluctuations in skill throughout longer surgical procedures. These findings help uncover different techniques of controlling for task time in surgical skill evaluation as well as uncovering different ways current evaluation techniques can be biased to speed.

7.1 Contributions

- An exploration of past surgical skill evaluation contributions by the research community, including techniques which have been promising and areas where speed can effect calculations.

- A machine learning algorithm capable of classifying surgical performances with state-of-the-art accuracy in a manner which controls for time of task completion using tooltip kinematic position data.

- An in-depth examination of speed perception defining the manner in which non-expert human raters are biased to perceive surgeons as performing better when they appear to be moving faster, in a manner which doesn't extend to novice surgeons.

- A deeper examination of a sub-set of highly skilled robotic surgeons, examining how this group of surgeons is perceived as even more proficient when they appear to be making quicker movements.

- A statistical analysis of velocity profile models for surgeon hand movements, finding that expert surgeons are *less* likely to follow these models in their movements.

- A dataset of over 340 de-identified and labeled robotic surgery cases which consist of high quality video as well as tooltip kinematic data recorded from da Vinci surgical robots.

- A study finding that performing a virtual reality warm-up module prior to robotic surgery did not significantly improve the skill of surgeons, as measured by motion metrics, non-expert crowds, and faculty surgeons.

- An analysis exploring the temporal fluctuations in surgical skill, questioning the standard reporting metrics in the field of surgical skill evaluation.

7.2 Limitations

There are a number of limitations to be noted from this work.

7.2.0.0.1 Speed Perception Studies The initial study which used dry lab laparoscopic task videos used only ten videos, with five surgeons in each of the two skill groups. It is possible that if more statistical power were obtained by using a dataset with a larger volume of novices, the negligible increase in scores at quicker playback speeds for novice surgeons would no longer appear.

7.2.0.0.2 Long Short-Term Memory Study First and foremost, these tasks were simulated procedures, and the proposed algorithms and techniques may perform differently on real surgeries. The proposed method also only analyzes tool motion data, which may not contain sufficient data required for complete skill classification. The results from the crowd reassessment could be in part due to differences in the user interface of the skill evaluation web pages used for the two different occurrences of testing.

7.2.0.0.3 Robotic Surgery Readiness Study The RSR project consisted of only robotic, minimally invasive surgeries. No laparoscopic or open surgeries were used. It is possible that skills needed for robotic surgery are fundamentally different than other surgical fields, causing no warm-up effect to be noticeable in this study. Only five warm-up modules were assessed as possible warm-ups to use for this study.

7.2.0.0.4 Rating Variability Study There was no semantic segmentation of these tasks, only temporal "chunking" into one minute-segments. Although videos of real surgeries were used, the dataset was fairly sparse, with only 12 videos of 15 minute surgeries. It is also not clear if the one minute duration used for segmentation is the optimal duration to use in obtaining crowd ratings. Shorter or longer segments of

surgery may be more beneficial for accurate skill evaluation. In this work, the one minute ratings were obtained 3 months after the full duration videos were rated. This may raise concerns about reliability (and comparability) of crowd ratings from studies separated by large time intervals.

7.3 Future of Surgical Skill Research

Several institutions are working towards developing new techniques to evaluate the skill of a surgeon. Extra care should be taken in the future to prevent creating techniques with biases towards faster surgeons, and instead use that data as one aspect of a broader list of attributes to evaluate. Making datasets open to the public for collaboration in skill evaluation research is the most beneficial way to innovate and build a safer surgical system, in which the annual number of medical errors is severely decreased. In the future, the creators of the datasets mentioned in this thesis plan to make them publicly available for other research institutions to use. Furthermore large professional societies of surgeons should be educated to use objective methods of evaluating skill, as opposed to subjective opinions. The collaboration of clinicians and scientists is a needed step forward to create a better future for health care.

References

[1] D. Holst, T. Kowalewki, L. White, T. Brand, J. Harper, M. Sorenson, S. Kirsch, and T. Lendvay. Crowd-sourced assessment of technical skills: An adjunct to urology resident surgical simulation training. *Journal of Endourology*, 29(5):604–609, 2015.

[2] A. Jacobs, J. Pinto, and M. Shiffrar. Experience, context, and the visual perception of human movement. *Journal of Experimental Psychology: Human Perception and Performance*, 30:822–835, 2004.

[3] T. Flash and N. Hogan. The coordination of arm movements an experimentally confirmed mathematical model. *The Journal of Neuroscience*, 5(7):1688–1703, 1985.

[4] M. Makary and M. Daniel. Medical error - the third leading cause of death in the us. *The BMJ*, 2016.

[5] L. Kohn, J. Corrigan, and M. Donaldson. To err is human. *National Academy Press*, 2000.

[6] J. Rosen, M. Salazzo, B. Hannaford, and M. Sinanan. *Medicine Meets Virtual Reality*, 2001.

[7] C.E. Reiley and G.D. Hager. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. *Medical Image Computing and Computer-Assisted Intervention*, pages 435–442, 2009.

[8] G. Megali, S. Sinigaglia, O. Tonet, and P. Dario. Modelling and evaluation of surgical performance using hidden markov models. *IEEE Transactions on Biomedical Engineering*, 53(10), 2006.

[9] M. Panzner and P. Cimiano. Comparing hidden markov models and long short term memory neural networks for learning action representations. *Semantic Computing Group*.

[10] H. I. Krebs, M.L. Aisen, B.T. Volpe, and N. Hogan. Quantization of continuous arm movements in humans with brain injury. *Proceedings of the National Academy of Sciences*, 96(2):4645–4649, April 1999.

[11] B. Rohrer and N. Hogan. Avoiding spurious submovement decompositions ii: a scattershot algorithm. *Biological Cybernetics*, 94:409–414, March 2006.

[12] C. Atkeson and J. Hollerbach. Kinematic features of unrestrained vertical arm movements. *Journal of Neuroscience*, 5(9):2318–2330, 1985.

[13] T.M. Kowalewski, L.W. White, T.S. Lendvay, I.S. Jiang, R.S. Sweet, A. Wright, B. Hannaford, and M.N. Sinanan. Beyond task time: automated measurements augments fundamentals of laparoscopic skills methodology. *Journal of Surgical Research*, 192(2):329–338, 2014.

[14] et. al. K. Ghani, B. Comstock. Technical skill assessment of surgeons performing robot-assisted radical prostatectomy: Relationship between crowdsourced review and patient outcomes. In *American Urological Association*, Boston, MA., 2017.

[15] L. Hammon and H. Hippner. Crowdsourcing. *Business and Information Systems Engineering*, 4(27):163–166, 2012.

[16] James Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[17] F. Galton. Vox populi. *Nature*, 75:450–451, 1907.

[18] A. Carvalho, S. Dimitrov, and K. Larson. How many crowdsourced workers should a requester hire. *Annals of Math and Artificial Intelligence*, 78(1):45–72, 2016.

[19] T.M. Kowalewski, B. Comstock, R. Sweet, C. Schaffhausen, A. Menhadji, T. Averch, G. Box, T. Brand, M. Ferrandino, J. Kaouk, B. Knudsen, J. Landman, B. Lee, B.F. Schwartz, E. McDougall, and T.S. Lendvay. Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. *The Journal of Urology*, 195(6):1859–1865, 2016.

[20] Inc. CSATS. Validation, 2019.

[21] L. White, T. Kowalewski, R. Dockter, B. Comstock, B. Hannaford, and T. Lendvay. Crowd-sourced assessment of technical skill (c-sats): A valid method for discriminating basic robotic surgery skills. *Journal of Endourology*, 29, 2015.

[22] C. Chen, L. White, T. Kowalewski, R. Aggarwat, C. Linnot, B. Comstock, K. Kuksenok, C. Aragon, D. Holst, and T. Lendvay. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of Surgical Research*, 187(1):65–71, 2014.

[23] N. Aghdasi, R. Bly, L. White, B. Hannaford, K. Moe, and T. Lendvay. Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *Journal of Surgical Research*, 196(2):302–306, 2015.

[24] M.C. Vassiliou, L.S. Feldman, C.G. Andrew, S. Bergman, K. Leffondre, D. Stanbridge, and G.M. Fried. A global assessment tool for evaluation of intraoperative laparoscopic skills. *The American Journal of Surgery*, 190(1):107–113, 2005.

[25] R. Plamondon, A. M. Alimi, P. Yergeau, and F. Leclerc. Modelling velocity profiles of rapid movements: a comparitive study. *Biological Cybernetics*, 69:119–128, January 1993.

[26] N.E. Berthier and R. Keen. Development of reaching in infancy. *Experimental Brain Research*, 169:507–518, December 2005.

[27] S. Balasbramanian, A. Melendez-Calderon, and E. Burdet. A robust and sensitive metric for quantifying movement smoothness. *IEEE Transactions of Biomedical Engineering*, 59(8):2126–2136, August 2012.

[28] S. Balasubramanian, A. Melendez-Calderon, A. Roby-Brami, and E. Burdet. On the analysis of movement smoothness. *Journal of NeuroEngineering and Rehabilitation*, 112, December 2015.

[29] A. Georgopoulos, J. Kalaska, and J. Massey. Spatial trajectories and reaction times of aimed movements: Effects of practice, uncertainty, and change in target location. *Journal of Neurophysiology*, 46(4):725–743, 1981.

[30] P. Morasso. Spatial control of arm movements. *Experimental Brain Research*, 42:223–227, 1981.

[31] S.W. Keele. Movement control in skilled motor performance. *Psychological Bulletin*, 70(6):387–403, 1968.

[32] W.S. McCulloch and W.H. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematic Biophysics*, 5:115–133, 1943.

[33] B. Widrow. Adaptive sampled-data systems — a statistical theory of adaptation. *IRE WESCON Convention Record*, 4:74–85, 1959.

[34] Simple model of a small artificiaal neural network, 2017.

[35] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[36] A. Graves, N. Jaitly, and A. Mohamed. Hybrid speech recognition with deep bidirectional lstm. *IEEE Workshop on Automatic Speed Recognition and Understanding*, 2013.

[37] H. Soltau, H. Liao, and H. Sak. Neural speech recognizer: Acoustin-to-word lstm model for large vocabulary speech recognition. *arXiv:1610.09975*, 2016.

[38] Siri Team. Deep learning for siri's voice: On-device deep mixture density networks for hybrid unit selection synthesis, 2017.

[39] J. Chen, P.J. Oh, N. Cheng, A. Shah, J. Montez, A. Jarc, L. Guo, I.S. Gill, and A.J. Hung. Use of automated performance metrics to measure surgeon performance during robotic vesicourethral anastomosis and methodical development of a training tutorial. *Journal of Urology*, 200(4):895–902, 2018.

[40] J.H. Nguyen, J. Chen, S.P. Mashall, S. Ghodoussipour, A. Chen, I.S. Gill, and A.J. Hung. Using objective robotic automated performance metrics and task-evoked pupillary response to distinguish surgeon expertise. *World Journal of Urology*, 2019.

[41] A. Hung, J. Chen, T. Nilanon, A. Jarc, M. Titus, P.J. Oh, I.S. Gill, and Y. Liu. Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. *Journal of Endourology*, 32(5), 2018.

[42] A. Zia, C. Zhang, X. Xiong, and A.M. Jarc. Temporal clustering of surgical activities in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 12(7):1171–1178, 2017.

[43] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, M.A. Clements, and I. Essa. Automated assessment of surgical skills using frequency analysis. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. 18th International Conference*, volume 9349, pages 430–438, Munich, Germany, September 2015.

[44] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition*, 3(15):1119–1125, November 1994.

[45] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[46] I. Laptev. On space-time interest points. *Internation Journal of Computer Vision*, 64(2):107–123, 2005.

[47] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, June 23-28 2008. IEEE Xplore, https://ieee.org. Date of Access: April 4, 2018.

[48] Fundamental of features and corners.

[49] Stanford University. Introduction to computer vision - tutorial 2: Image matching.

[50] Y. Li, R. Xia, Q. Huang, W. Xie, and X. Li. Survey of spatio-temporal interest point detection algorithms in video. *IEEE Access*, 5:10323–10331, 6 2017.

[51] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. *British Machine Vision Conference*, September 2009.

[52] Y. Sharma, V. Bettadapura, N. Hammerla, S. Mellor, R. Mcnaney, P. Olivier, E. Deshmukh, A. Mccaskie, and I. Essa. Video based assessment of osats using sequential motion textures. *Georgia Institute of Technology Library*, 2014.

[53] A. Zia and I. Essa. Automated surgical skill assessment in rmis training. *International Journal of Computer Assisted Radiology and Surgery*, 13(5):731–739, March 2018.

[54] A. Zia, C. Zhang, X. Xiong, and A.M. Jarc. Temporal clustering of surgical activities in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 12(7):1171–1178, July 2017.

[55] Y. Sharma, V. Bettadapura, T. Plotz, N. Hammerla, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa. Video based assessment of osats using sequential motion textures. In *Fifth Workshop on Modeling and Monitoring of Computer Assisted Intervations – M2CAI*, 2014.

[56] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, and I. Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. 2017.

[57] M.J. Fard, S. Ameri, E.R. Darin, R.B. Chinnam, A.K. Pandya, and M.D. Klein. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. *International Journal of Medical Robotics*, 14(1), 2018.

[58] I. Funke, S.T. Mees, J. Weitz, and S. Speidel. Video-based surgical skill assessment using 3d convolutionaal neural networks. 2019.

[59] Z. Wang and A.M. Fey. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. 2019.

[60] L.T. Kozlowski and J.E. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics*, 21:575–580, 1977.

[61] P. Veto, W. Einhäuser, and N.F. Troje. Biological motion distorts size perception. *Scientific Reports*, 7(42576):1–6, 2017.

[62] C. D Barclay, J.E. Cutting, and L.T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics*, 23(2):145–152, 1978.

[63] R. Blake and M. Shiffrar. Perception of human motion. *Reviews in Advance*, 15(8), 2007.

[64] J.A. Beintema, K. Georg, and M. Lappe. Perception of biological motion from limited-lifetime stimuli. *Perception and Psychophysics*, 68:613–624, 2006.

[65] M.A. Giese and M. Lappe. Measurement of generalization fields for the recognition of biological motion. *Vision Research*, 42:1847–1858, 2002.

[66] D. Jokisch and N.F. Troje. Biological motion as a cue for the perception of size. *The Journal of Vision*, 3(4):252–264, 2003.

[67] J. Bruin. Introduction to linear mixed models.

[68] Newsom. Distinguishing between random and fixed: Variables, effects, and coefficients, 2019.

[69] J.D. Kelly, A. Petersen, T.S. Lendvay, and T.M. Kowalewski. Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. *International Journal of Computer Assisted Radiology and Surgery*, SUBMITTED.

[70] R. Dockter, T.S. Lendvay, R.M. Sweet, and T.M. Kowalewski. The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. *International Journal of Computer-Assisted Radiology and Surgery*, 12:1151–1159, 2017.

[71] H.C. Lin, I. Shafran, T.E. Murphy, A.M. Okamura, D.D. Yuh, and G.D. Hager. Automatic detection and segmentation of robot-assisted surgical motions. *Medical Image Computing and Computer-Assisted Intervention*, 3749, 2005.

[72] J.D. Birkmeyer, J.F. Finks, A. O'Reilly, M. Oerline, A.M. Carlin, A.R. Nunn, J. Dimick, M. Banerjee, and N.J. Birkmeyer. Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, 369(15):1434–1442, 2013.

[73] J.D. Kelly, A. Petersen, T.S. Lendvay, and T.M. Kowalewski. The effect of video playback speed on surgeon technical skill perception. *International Journal of Computer Assisted Radiology and Surgery*, 2020.

[74] M. Schuster and K.P. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(5), 1997.

[75] A.M. Derossis, G.M. Fried, M. Abrahamowicz, H.H. Sigman, J.S. Barkun, and J.L. Meakins. Development of a model for training and evaluation of laparoscopic skills. *American Journal of Surgery*, 175:482, 1998.

[76] G.M. Fried. Fls assessment of competency using simulated laparsocopic tasks. *Journal of Gasteroenterology Surgery*, 12:210, 2008.

[77] J.H. Peters, G.M. Fried, L.L. Swanstrom, N.J. Soper, L.F. Silin, B. Schirmer, and K. Hoffman. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery*, 135:21.

[78] R.M. Seete, R. Beach, F. Sainfort, P. Gupta, T. Reihsen, L.H. Poniatowski, and E.M. McDougall. Introduction and validation of the american urological association basic laparoscopic urology surgery skill curriculum. *Journal of Endourology*, 26:190.

[79] T.M. Kowalewski, R. Seete, T.S. Lendvay, A. Menhadji, T. Averch, G. Box, T. Brand, M. Ferrandino, J. Kaouk, B. Knudsen, J. Landman, B. Lee, B.F. Schwartz, and E. McDougall. Validation of the aua blus tasks. *Journal of Urology*, 195:998, 2016.

[80] A. French, K. Seidel, T.S. Lendvay, and T.M. Kowalewski. Role of contextual information in skill evaluation of minimally invasive surgical training procedures. *Hamlyn Symposium on Medical Robotics*, 2018.

[81] J.D. Kelly, Nicholas Heller, A. Petersen, T.S. Lendvay, and T.M. Kowalewski. The effect of video playback speed on robotic surgery skill evaluation. *International Journal of Computer Assisted Radiology and Surgery*, SUBMITTED.

[82] A.A. Gumbs, N.J. Hogle, and D.L. Fowler. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *Journal of the American College of Surgeons*, 204(2):308–313, 2007.

[83] A. Bajcsy, D.P. Losey amd M.K. O'Malley, and A.D. Dragan. Learning from physical human corrections, one feature at a time. *In Proceedings of 2018 ACM/IEEE International Conference on Human-Robot Interaction.*

[84] T.M. Kowalewski. Real-time quantitative assessment of surgical skill. *PhD Thesis, University of Washington*, 2012.

[85] FFMPEG Developers, version 4.1.3 [software] edition, 2016.

[86] P. Holoborodko. Smooth noise robust differentiators. `http://www.holoborodko.com/pavel/numerical-methods/numerical-derivative/smooth-low-noise-differentiators/`, year = 2008.

[87] Python Software Foundation. *Python Language Reference*. `http://www.python.org`, version 3.6 [software] edition.

[88] *MATLAB R2019a [Software]*. The Mathworks Inc., Natick, Massachusetts.

[89] F. C. Goh, D.W. Goldfarb, J.C. Sander, B.J. Miles, and B.J. Dunkin. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. *The Journal of Urology*, 187(1):247–252, 2012.

[90] R.M. Satava, A.G. Gallagher, and C.A. Pellegrini. Surgical competence and surgical proficiency: definitions, taxonomy, and metrics. *Journal of the American College of Surgeons*, 196(6):933–7, 2003.

[91] A. French, T.S. Lendvay, R.M. Sweet, and T.M. Kowalewski. Predicting surgical skill from the first n seconds of a task: Value over task time using the isogony

principle. *The International Journal of Computer Assisted Radiology and Surgery*, 12(7):1161–1170.

[92] R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.R-project.org/, Vienna, Austria, 2018 edition.

[93] J.D. Kelly, A. French, M. Nash, L. Meryman, N. Heller, N. Organ, E. George, R. Smith, M. Sorensen, T. Brand, B. Comstock, T.M. Kowalewski, and T.S. Lendvay. Robotic surgery readiness: A randomized controlled study of virtual reality warm-up prior to robot-assisted surgery. *Journal of the American College of Surgeons*, In Preparation.

[94] C. Zhan and M. Miller. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. *JAMA*.

[95] S. Murhpy, J. Xu, and K. Kochanek. Preliminary data for 2010. *National Vital Statistics Reports*, 60(4), 2012.

[96] J. Shreve. The economic measurement of medical errors. *Society of Actuaries*, 2010.

[97] G. Sroka, L. Feldman, and M. Vassiliou. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room - a randomized controlled trial. *American Journal of Surgery*, 199:115, 2010.

[98] D. Stefanidis, J.R. Korndorffer Jr., R. Sierra, C. Touchard, J.B. Dunne, and D.J. Scott. Skill retention following proficiency-based laparoscopic simulator training. *Surgery*, 138:165–170, 2005.

[99] R.S. Perez, A. Skinner, P. Weyhrauch, J. Nehaus, C. Lathan, S.D. Schwaitzberg, and C.G. Cao. Prevention of surgical skills decay. *Military Medicine*, 178:588–592, 2013.

[100] K.H. Sheetz, J. Claflin, and J.B. Dimick. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA Network Open*, 3(1), 2020.

[101] A.B. Haynes, T.G. Weiser, W.R. Berry, S.R. Lipsitz, A.-H.S. Breizat, E.P. Dellinger, T. Herbosa, S. Joseph, P.L. Kibatala, M.C.M. Lapitan, A.F. Merry, K. Moorthy, R.F. Reznick, B. Taylor, and B.B. Gawande. A surgical safety checklist to reduce morbidity and mortality in a global population. *NEJM*, 360:491–499, 2009.

[102] J. Nascon and R.A. Schmidt. The activity-set hypothesis for warm-up decrement. *Journal of Motor Behavior*, 3:1–16, 1971.

[103] C.A. Wrisberg, A.W. Salmoni, and R.A. Schmidt. Warm-up effects in the learning of discrete motor skills. *Acta Psychologica*, 29:311–320, 1975.

[104] M.A. Anshel. The effect of arousal on warm-up decrement. *Research Quarterly for Exercise and Sport*, 56:1–9, 1985.

[105] M.H. Anshel and C.A. Wrisberg. Reducing warm-up decrement in the performance of the tennis serve. *J Sport and Exercise Psychology*, 15:290–303, 1993.

[106] P.A. Wetter. Resubmission of gap analysis workshop for training for reintegration of surgical skills. *Proceedings of the 2011 Society of Laparoendoscopic Surgeons Conference*, 2011.

[107] W. Arthur Jr., B. Winston Jr., P.L. Stanush, and T.L. McNelly. Factors that influence skill decay and retention: A quantitative review and analysis. *Human Performance*, 11:57–101, 1998.

[108] N.E. Seymour, A.G. Gallagher, S.A. Roman, M.K. OBrien, V.K. Bansal, D.K. Andersen, and R.M. Satava. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg*, 236:458–463, 2002.

[109] K. Kahol, R. Satava, J. Ferrara, and M.L. Smith. Effect of short-term pretrial practice on surgical proficiency in simulated environments: a randomized trial of the preoperative warm-up effect. *JACS*, 208:255, 2009.

[110] D. Calatayud, S. Arora, R. Aggarwal, I. Kruglikova, S. Schulze, P. Funch-Jensen, and T. Grantcharov. Warm-up in a virtual reality environment improves performance in the operating room. *Ann Surg*, 251:1181, 2010.

[111] T.S. Lendvay, T.C. Brand, L. White, T. Kowalewski, S. Jonnadula, L. Mercer, D. Khorsand, J. Andros, B. Hannaford, and R.M. Satava. Virtual reality robotics surgery warm-up improves surgical performance: A prospective randomized controlled study. *JACS*, 216:1181–1192, 2013.

[112] D. Stefanidis, J.R. Korndorffer Jr., R. Sierra, C. Touchard, J.B. Dunne, and D.J. Scott. Skill retention following proficiency-based laparoscopic simulator training. *Surgery*, 138:165–170, 2005.

[113] Inc. Mimic Technologies. da vinci skills simulator (dvss), 2018.

[114] Y. Sharon, T.S. Lendvay, and I. Nisky. Instrument orientation metrics for robot-assisted and open surgical skill evaluation. *IEEE Transactions on Human-Machine Systems*, 2017.

[115] A.J. Hung, P. Zehnder, M.B. Patil, J. Cai, C.K. Ng, M. Aron, I.S. Gill, and M.M. Desai. Face, content and construct validity of a novel robotic surgery simulator. *Journal of Urology*, 186:1019–1025, 2011.

[116] J.B. Bakdash and L.R. Marusich. Repeated measures correlation, 2017.

[117] J.D. Kelly, M. Nash, N. Heller, T.S. Lendvay, and T.M. Kowalewski. Temporal variability of surgical technical skill perception in real robotic surgery. *International Journal of Computer Assisted Radiology and Surgery*, SUBMITTED.

[118] N. Geri, A. Winer, and B. Zaks. Challenging the six-minute myth of online video lectures: Can interactivity expand the attention of learners? *Online Journal of Applied Knowledge Management*, 5(1):101–111, 2017.

[119] M. Schmitt, D.C.A. Bulterman, and P.S. Cesar. The contrast effect: Qoe of mixed-video qualities at the same time. *Quality and User Experience*, 3(7), 2018.

[120] Inc. C-SATS, 2020.

[121] H.W.R. Schreuder. Virtual reality training for robotic surgery. In *Hospital Healthcare Europe*. Cogora Limited, 2014.

[122] T.S. Lendvay, T. Brand, A. French, L. Meryman, N. Organ, M. Sorensen, and T.M. Kowalewski. *Society of Pediatric Urology, Fall Congress*, 2018.

[123] A. Gawande. The number of surgical procedures in an american lifetime in 3 states. *Journal of the American College of Surgery*, 5(7):1688–1703, 1985.

[124] A. Zia, Y. Sharma, V. Bettadapura, E.L. Sarin, T. Ploetz, M.A. Clements, and I. Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International Journal of Computer Assisted Radiology and Surgery*, 11(9):1623–1636, September 2016.

[125] Y. Sharma, T. Plotz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa. Automated surgical osats prediction from videos. In *2014 IEEE 11th International Symposium on Biomedical Imaging – ISBI*, Beijing, China, April 29–May 2 2014. IEEE Xplore.

[126] A. Peuna, J. Hekkala, M. Haapea, J. Podlipska, A. Guermazi, S. Saarakkala, M. Nieminen, and E. Lammentausta. Variable angle gray level co-occurrence matrix analysis of t2 relaxation time maps reveals degenerative changes of cartilage in knee osteoarthritis: Oulu knee osteoarthritis study. *Journal of Magnetic Resonance Imaging*, 47:1316–1327, 2018.

[127] G. Zhenhua, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Transactions on Image Processing*, 19(6):1657–1663, 2010.

[128] J. Shi, X. Weng, S. He, and Y. Jiang. Biological motion cues trigger reflexive attentional orienting. *Cognition*, 117(3):348–354, 2010.

[129] S. Gharghabi, Y. Ding, C.M. Yeh, K. Kamgar L. Ulanova, and E. Keogh. Matrix profile viii domain agnostic online semantic segmentation at superhuman performance levels. In *2017 IEEE International Conference on Data Mining*, New Orleans, LA, Nov 18–Nov 21 2017. IEEE Xplore.

Appendix A

Robotic Surgery Readiness Supplemental Figures

The following supplemental figures are added, pertaining to Chapter 6, from the RSR study. Figure A.1 is a robotic surgery video from the Si platform after being edited to remove personally identifiable information of the performing surgeon. Figure A.2 are plots of the kinematic tooltip positions from a sample video of robotic surgery. A screenshot of the website created to be used for obtaining faculty surgeon video reviews is shown in Figure A.3. Figure A.4 displays cumulative data from several different tasks combined of both the 3-D coordinates as well as position histograms in each coordinate plane, to give an idea for the general movements occurring during robotic surgery tasks.

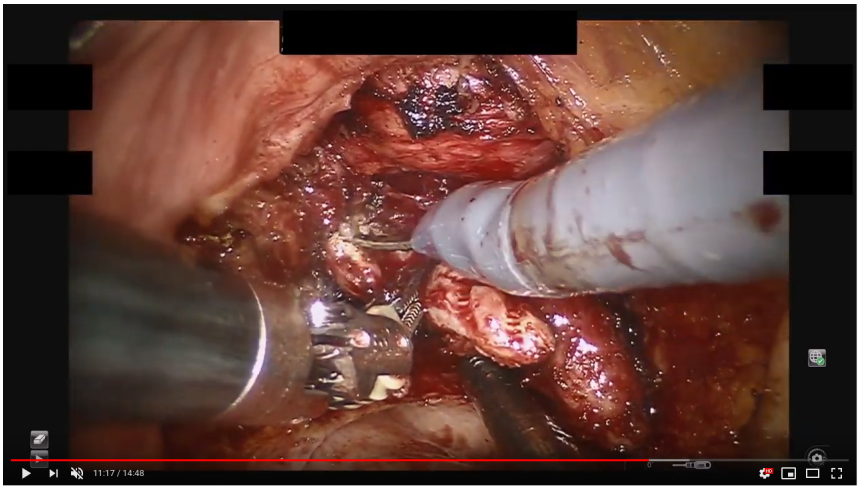


Figure A.1: Video of an Si daVinci surgery, after editing the video to the correct times and placing black boxes over the identifying pieces of information.

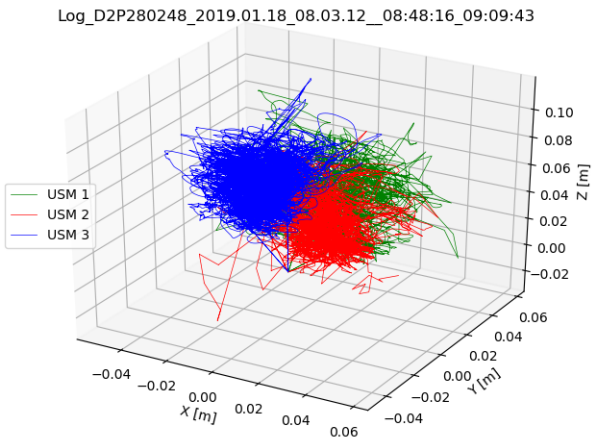


Figure A.2: 3-D Plot of the kinematic tooltips from a robotic surgery which corresponds with video edited at the same time range. USM 1-3 refers to the Universal Serial Manipulators, the surgical arms.

Figure A.3: A view of the website faculty surgeons would see when visiting to evaluate robotic surgeon technical skill.

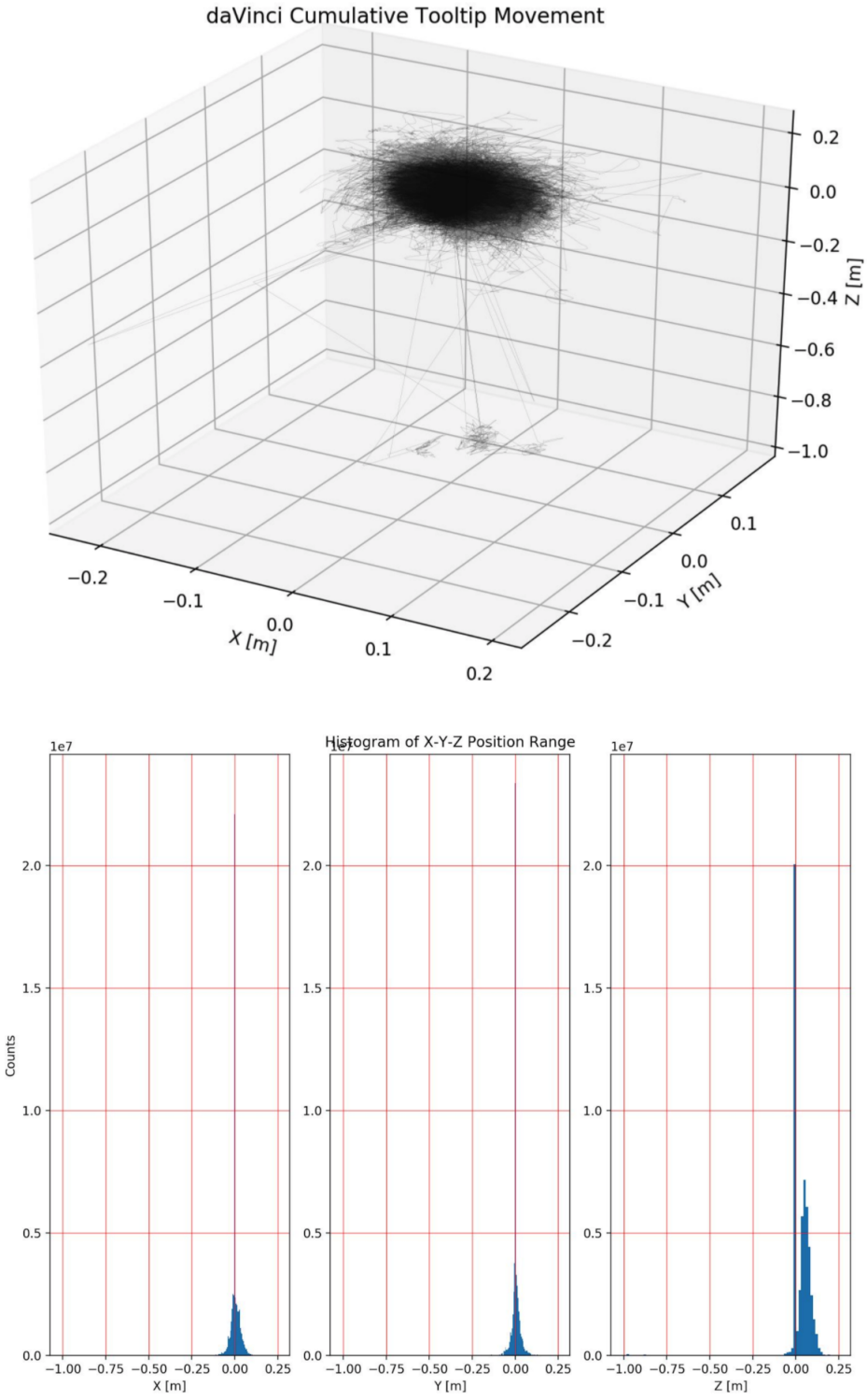


Figure A.4: Cumulative positions of several combined tasks from the RSR study, in both 3-D coordinates, as well as a histogram for frequency of position in each axis.

A.0.1 Kinematic Metrics

Several performance metrics were used to attempt to rank and visualize the relative skill of each performance. Some of these included rate of orientation change, normalized angular displacement [114], as well as spectral arc length (Chapter 4). Normalized angular displacement is defined as

$$A = \frac{1}{P} \sum_{j=1}^{N-1} |\delta\theta_{j,j+1}| \tag{A.1}$$

in which P is the path length, or the length of movement during a continuous movement between stops. This metric is a score of the overall change in orientation during a movement, which could also be thought of as the angular path. The rate of orientation change is

$$C = \frac{1}{N-1} \sum_{j=1}^{N-1} \frac{\delta\theta_{j,j+1}}{\delta t_{j,j+1}} \tag{A.2}$$

where $t_{j,j+1}$ is the time difference between samples. Other computed metrics were mean velocity, median velocity, maximum velocity, minimum velocity, total time, path length, economy of motion, and acceleration.

A.0.2 Event Metrics

The da Vinci software records various event measures which have been shown in the past to correlate to skill. Of these, the metrics recorded for each case are shown in Table A.1.

Event	Description
ENERGY_CTRL	Surgeon consoles are switched
CAMCTRL_BTN	Camera is moved
ARM_SWAP	Arms are switched between one another
HEAD_SENSE	Surgeon presses head in or out of console
PORT_CLUTCH	Surgeon presses clutch to move instruments

Table A.1: Recorded event measures for the RSR study.