AWARD NUMBER: W81XWH-15-2-0030

TITLE: Robotic Surgery Readiness (RSR): A Prospective Randomized Skills Decay Recognition and Prevention Study

PRINCIPAL INVESTIGATOR: Thomas Lendvay

CONTRACTING ORGANIZATION: University of Washington

Seattle, WA 98195

REPORT DATE: August 2017

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
XX August 2017	Annual	15 July 2016- XXXXXXXXXX 14 July 2017

Robotic Surgery Readiness (RSR): A Prospective Randomized Skills Recognition and Prevention Study

5a. CONTRACT NUMBER	W81XWH-15-2-0030
5b. GRANT NUMBER	12195004
5c. PROGRAM ELEMENT NUMBER	
5d. PROJECT NUMBER	
5e. TASK NUMBER	
5f. WORK UNIT NUMBER	

6. AUTHOR(S)

Thomas Lendvay, MD

E-Mail: Thomas.lendvay@seattlechildrens.org

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Washington
4333 Brooklyn Ave NE
Seattle, WA 98195-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

0

14. ABSTRACT

Preliminary research has demonstrated that surgical warm-up, either on a virtual reality (VR) simulator or reality-based module, can improve surgical performance, yet metrics to identify and counteract skills decay are not readily available to provide targeted curricula. We will establish performance signatures of Robotic Surgery Readiness (RSR) through tasks on the da Vinci robotic virtual reality simulator, testing the role intervals of inactivity have on task performance. These will be used to develop a simulation curriculum that brings the inactive surgeon to RSR.

AIM 1: We are nearing the end of recruitment for AIM 1 with 28/40 completed subjects. Another 8 subjects have reached proficiency and are in their trial sessions. The recruitment has taken longer than anticipated due to challenging clinic rotations schedules and deployments for some subjects. Data analysis cannot commence until all 40 subjects are complete, yet data upload standardization processes are established.

AIM 2: We have settled on the most optimal method for extracting kinematic and video data from the da Vinci robots for Aim 2 using the Intuitive dVLogger system which captures these data directly from the API of the robot, thus standardizing the method for capture. The work flow has been tested at two sites and is being optimized. We have not recruited subjects for AIM 2, yet, pending completion of recruitment in Aim 1.

Data Analysis: Pending completion of Aim 1 trial sessions.

15. SUBJECT TERMS

Robotic Surgery, Readiness, da Vinci Simulator, Virtual Reality, Simulation Curriculum, GEARS – Global Evaluation Assessment of Robotic Skills, Surgical Education

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	44	19b. TELEPHONE NUMBER *(Include area code)*

Table of Contents

Page

1. INTRODUCTION:

We will establish the performance signatures of Robotic Surgery Readiness (RSR) through tasks on the da Vinci robotic virtual reality simulator by testing the role intervals of inactivity have on task performance. These signatures will be used to develop a simulation curriculum that brings the inactive surgeon to RSR. The curriculum effectiveness will be tested in the operating room on practicing surgeons performing patient surgery with and without the RSR warm-up curriculum. We will enroll surgical residents and faculty for hypothesis testing. Objective technical performance and Global Evaluative Assessment of Robotic Skills (GEARS) scoring will be correlated by the Principal Investigator (Dr. Thomas Lendvay - UW) and Co-Investigator (Dr. Timothy Kowalewski - UMN). Optimal methods for extracting surgeon performance metrics from the da Vinci Application Programming Interface (API) will be evaluated and developed through collaboration with the Intuitive Surgical Consultant (Simon DiMaio, Senior Research Manager). We will deliver practical, automated RSR assessment methods and a warm-up curriculum able to bring a robotic surgeon to his/her optimal state of readiness before patient surgery.

2. KEYWORDS: *Provide a brief list of keywords (limit to 20 words).*

Robotic Surgery
Readiness
da Vinci Simulator
Virtual Reality
Simulation Curriculum
GEARS - Global Evaluative Assessment of Robotic Skills
Surgical Education

3. ACCOMPLISHMENTS:

What were the major goals of the project?

YEAR 1 (0-12 Months)

1) Study design and skill decay model construction, supplies purchasing and acquisition. *Completion date 9/30/2015, major supplies purchased thru 6/30/2016*
2) Set-up and flow within and between simulation centers.
   Completion date 9/30/2015
3) Development of robust methods for collecting, merging and verifying simulator, video and optical tracking data.
   Completion date 12/31/2015. Continued checks as more equipment comes online.
   Tool motion metric capturing technology development – ongoing.
4) Subject recruitment.
   Recruitment Continues at all 4 sites. 87% complete
5) Skills decay testing. *In progress*
6) Independent video review of VR simulator criterion performances using GEARS tool. *Not started.*
7) Analysis of performance metrics. *Not started.*

Deliverables: Quantifiable performance signatures of robotic surgery skills decay assessment. Initial analysis of data. Preliminary RSR warm-up curriculum.

YEAR 2 (12-24 Months)

1) Finalize and validate RSR curriculum and benchmarks. *Not started yet – recruitment and skills decay testing sessions are still underway.*
2) Intra-operative RSR warm-up subject recruitment. *Not started yet – recruitment and skills decay testing sessions still underway.*
3) RSR curriculum hypothesis testing, intra-operative data collection. *Not started yet – recruitment and skills decay testing sessions still underway.*
4) Independent video review of surgical performances using GEARS. *Not started yet – recruitment and skills decay testing sessions still underway.*
5) Testing of kinematic & video capture systems *In progress - started 4/2017.*
6) Building and refining the Aim 2 REDCap database *In progress – started 6/2017.*

Deliverable: Finalized RSR warm-up curriculum, initial dataset and data quality assessment.

YEAR 3 (24-36 Months)

1) Continued intra-operative RSR curriculum hypothesis testing.
2) Continued independent video review of operative performances using GEARS.
3) Biostatistical analysis and model cross-validation.
4) Abstract and Manuscript drafting.

Final Deliverables: Completed, validated RSR warm-up curriculum and assessment tools. Methodology for quantifying robotic surgery skills decay. Peer-reviewed publication, presentation at national meeting.

What was accomplished under these goals?

YEAR 2

Study design and skill decay model construction, supplies purchasing and acquisition.

Computers have been purchased and setup for all sites (UW, MAMC, VA, FL Hospital)

Set-up and flow within and between simulation centers.

Recruited subjects complete the intake demographics questionnaire and begin proficiency training. Subject details are kept at each site and only de-identified data is collected by the team at Minnesota (UMN). All subjects are given a unique identifier based on their location.

UMN provided early drafts and input on revisions for the data collection forms and a UW biostatistician continues to improve the REDCap data collection process. REDCap is now being at all sites to collect data and to randomize the subjects.

UMN has reworked the software to provide high definition video acquisition and compression and has successfully installed and monitored the acquisition of data from all sites. The data is synchronizing with our central database as designed. UNM continues to rectify the Google Drive videos, REDCap data, and dVLogger data to ensure all logs are complete and no data is lost and where necessary, investigates incomplete logs and proposes fixes.

A collaborative triparty agreement is now in place, brokered between the Office of Sponsored Research at the University of Washington, UMN and Intuitive Surgical, Inc. to provide kinematic data directly from the da Vinci robots.

Subject recruitment. *(Continues at all sites)*

Aim 1 Recruitment is underway at all four sites with the consented and proficient subjects.

Florida Hospital and UW successfully reconciled contract language which had delayed their active participation.

Skills decay testing. *(Continues at all sites)*

As shown in the table 1 below, 28 subjects have completed their study sessions.
MAMC has consented 35 subjects and 20 have progressed to completion. Five of continue to progress through their sessions and 8 are still working to achieve proficiency and 1 withdrew. One has been lost to follow-up.
UW/VA has consented 28 subjects. Ten subjects have been randomized and 8 of those have completed all study sessions. Two subjects are continuing through study sessions. Nine are working on proficiency, 3 have dropped out due to scheduling issues and 6 are lost to follow-up.
Florida Hospital has 4 consented; 1 has completed, 2 withdrew due to time constraints and another was relocated to another hospital.

Robotic Surgery Readiness (RSR) : Enrollment Report

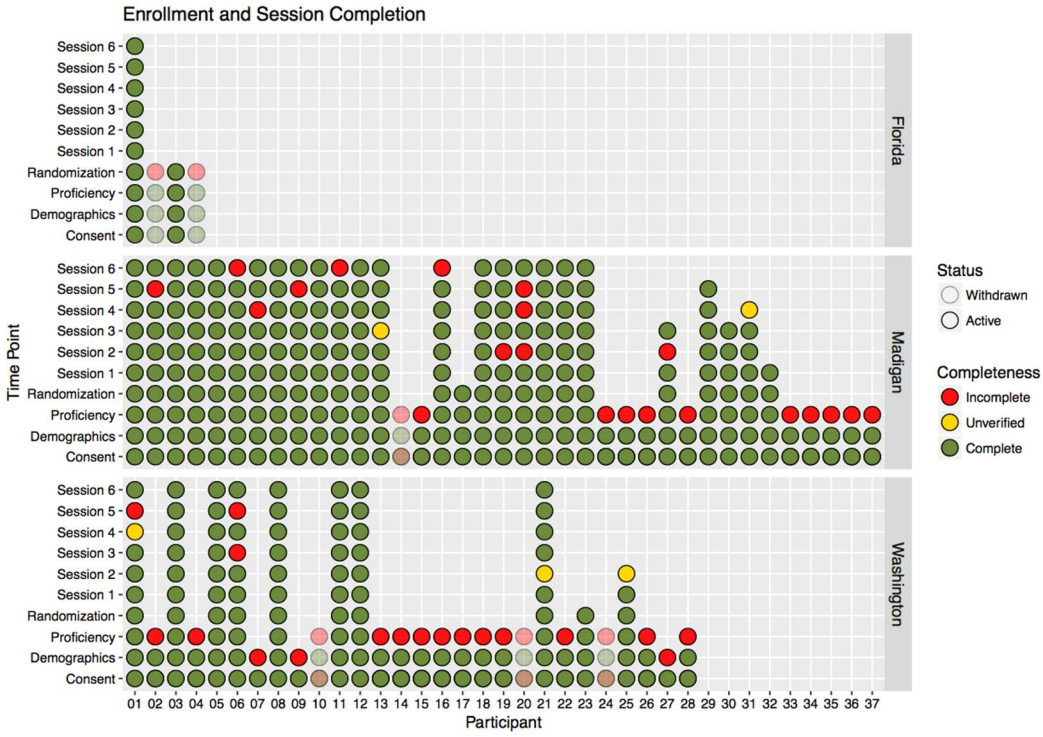
July 31, 2017

Table 1: Recruitment data summary by site, n (%)

	Florida	Madigan	Washington	Total
Target Enrollment	20	20	30	70
Total Randomized	**2 (10)**	**26 (130)**	**10 (33)**	**38 (54)**
Completed Participation	**1 (5)**	**20 (100)**	**8 (27)**	**29 (41)**
Consented*	4 (20)	37 (185)	28 (93)	69 (99)
Gender				
Female	1 (25)	9 (24)	8 (29)	18 (26)
Male	3 (75)	28 (76)	17 (61)	48 (70)
Other	0 (0)	0 (0)	0 (0)	0 (0)
Seniority				
Attending	4 (100)	11 (30)	2 (7)	17 (25)
Senior	0 (0)	9 (24)	9 (32)	18 (26)
Junior	0 (0)	17 (46)	14 (50)	31 (45)
Specialty				
Urology	0 (0)	12 (32)	11 (39)	23 (33)
Gynocology	1 (25)	8 (22)	2 (7)	11 (16)
General	2 (50)	8 (22)	10 (36)	20 (29)
Thoracic	0 (0)	0 (0)	0 (0)	0 (0)
Cardiothoracic	0 (0)	0 (0)	0 (0)	0 (0)
ENT	1 (25)	9 (24)	2 (7)	12 (17)
Handedness				
Left	1 (25)	1 (3)	1 (4)	3 (4)
Right	2 (50)	34 (92)	22 (79)	58 (84)
Ambidextrous	1 (25)	2 (5)	2 (7)	5 (7)
Musical Experience	2 (50)	22 (59)	18 (64)	42 (61)
Deployed	0 (0)	6 (16)	0 (0)	6 (9)

*Denominator for percentages below

1



Data entry is ongoing and REDCap forms are updated as needed for optimal data capture. Syncing issues were identified between the Florida and UMN and hardware at Florida could not be debugged remotely. The laptop was returned, debugged and mailed back to the Florida site.

A data logging time stamp issue was identified from the dVSS . Software was created to parse data from the logger into human readable format for rectification and logging purposes. Code was modified to meet end user needs and software was subsequently created to rectify DV logger data to REDCap logs.

We developed data collection tools for Aim 2. These tools, to be implemented on paper or through REDCap, include a participant demographics questionnaire, session data caption, and outcome data collection.

Designed and built the REDCap database for Aim 2. This included developing a randomization schema and system to randomize each participant to the RSR assignment

Site Visits

We conducted site visits to FL Hospital and to MAMC last October. These visits were very informative. FL Hospital has a wonderful facility but its distance from the hospital and subjects has been a problem. The FL Hospital team is very capable and continues to make efforts with the local surgery population to recruit them into the study. The MAMC coordinator has exceeded recruitment expectations for her site.

In April, team members met for 3 days in Seattle for orientation to and installation of the Intuitive DV Logger. Dr. Lendvay and Lois Meryman from UW, Dr. Kowalewski and Anna French from UM, and Evelyn George from Madigan and were in attendance. Omid Mohareri from Intuitive Surgical demonstrated the data capture process. We tested integration of dVLogger with our hardware (only Apple wireless worked without flaw, hence move to Apple devices for Aim 2 data collection/control)

Test data was captured, reviewed by the team and verified that full data streams are available from dVLogger for extracting sample metrics like path length and working volume (see sample plot below, Figure 1)

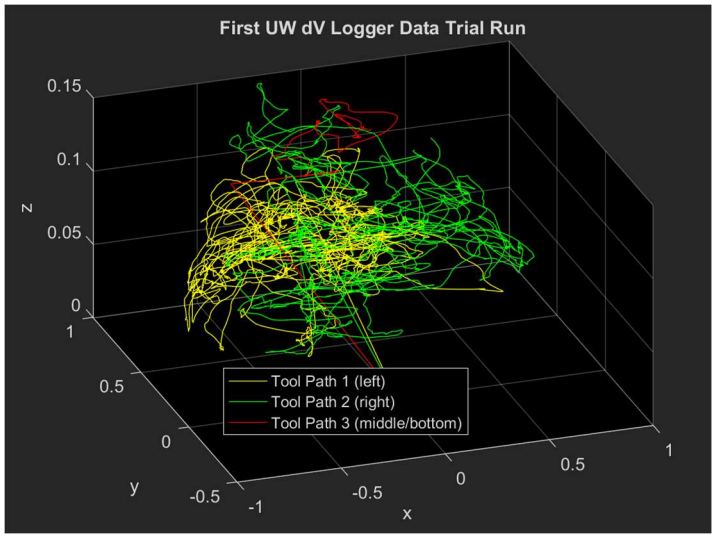


Figure 1. Sample plot

Protocol was developed for the implementation of Aim 2 and resulted in documentation and instructions for site administrators (see BEFORE case checklist below).

BEFORE Case checklist:
Ethernet cable connected
L and R SDI cables connected
Power cables connected
Confirm stereo configurations on tower
Confirm USB connection between hard drive and dV-Logger
Power on dV-Logger well before surgeon begins operating
Log into web-UI
Access webpage via http://10.42.0.1
Check hard drive storage space
Confirm that data is streaming
Confirm that recording has started BEFORE the procedure begins
If any items from the checklist were NOT addressed, please describe why:

What opportunities for training and professional development has the project provided?

All Aim 1 participants are provided training in robotic surgery simulation activities to meet proficiency. This has been accomplished through peer and one-on-one training with an expert. The Proficiency Training introduces novices and hones experienced clinicians in robotic object transfer, suturing, management of the third working arm, camera and instrument clutching skills. We have not provided "Professional development" opportunities.

How were the results disseminated to communities of interest?

Nothing to report.

What do you plan to do during the next reporting period to accomplish the goals?

YEAR 3

A site visit to MAMC is planned for August 1 to test the DV logger and to capture data from two surgeries in preparation for Aim 2. A Mobile App interface between the DV Logger and a mobile device that assures the system is operating properly will be tested.
Subject recruitment is expected to be completed and all subjects randomized. Recruitment for Aim 2 will commence.
Data acquisition is continually monitored by the UMN team.
The entire team is diligently working to ensure all data collection systems are in place for Aim 2 implementation. Refinements will continue for Aim 2 methods for collecting, merging and verifying simulator and video.
The Florida Hospital team is currently in discussions with the lead robotic urologist to schedule fellows for Aim 1.

4. IMPACT:

What was the impact on the development of the principal discipline(s) of the project?

We have developed a video and data capture system that allows remote software updates on each site's computer. This has minimized the need for any on-site software/hardware servicing. Furthermore, a workable user-interface was developed so that each site's coordinators can seamlessly capture video and upload data.

What was the impact on other disciplines?

A method for reliable seamless video capture, data tagging, and storage has a universal application in any training and skills assessment programs.

What was the impact on technology transfer?

Nothing to report.

What was the impact on society beyond science and technology?

Nothing to report.

5. CHANGES/PROBLEMS:

Changes in approach and reasons for change

Nothing to report

Actual or anticipated problems or delays and actions or plans to resolve them

As this is a multisite project, problems were anticipated. Due to the nature of funding a military site and the need for all funds to be used within that fiscal year we had to bring back the money allocated to MAMC and fund any manpower hours required centrally. The funding contract with Florida Hospital took longer than anticipated but is now in place.

Enrollment Barriers

Data collection is always the most unpredictable factor in a study. Enrollment has been less robust than anticipated.

While one site (MAMC) has proven exceptional, enrollment at UW+VA and Florida are lower than projected. The principal barrier appears to be subject inconvenience. Slower enrollment at UW/VA sites were due to staffing issues which have been resolved by the addition of dedicated staff
MAMC's simulator is in a room beside the clinic, near the physicians' offices, and is used in surgery one to two days per week. The subjects are in a relatively private space, with nearly unlimited access to the equipment, and are not required to access the operating suites or change into scrubs.
In comparison, the UWMC simulator is located in an operating room (OR), and is not available during surgery hours (i.e., ~07:00-19:00 weekdays). All users are required to change into scrubs.
The VA's simulator, when not in use, is packed in a storage closet adjacent to the OR; access to the equipment is dependent on the OR's usage. All users are required to change into scrubs. Further, the General Surgery residents share a key-card, limiting access to the locker rooms and OR.
In addition, the General Surgery residents in the UW/VA program are required to work a minimum of 80 hours per week. The General Surgery Director has determined that time related to this study may not be included in that commitment, and must be scheduled separately. This restriction has been a deterrent to proceeding through the study.
At FL Hospital, the simulator facility is approximately 20 minutes' drive from the hospital proper. Subjects are understandably reluctant to invest in the additional commute.
Randomization of consented subjects has been impacted by clinical rotations and busy schedules, limiting residents to participate once they have consented to the study.

Changes that had a significant impact on expenditures

Since Florida has not been as successful at recruitment and retention, more resources are needed at University of Washington/VA to cover the additional FTE for Year 3. The DV Logger will be able to extract the necessary kinematic data from the daVinci Robot, and thus we hope to seek approval from the DoD to repurpose monies allocated to the optical tracker to personnel FTE, especially at UW and UMN. A formal budget modification request will be forthcoming in Year 3 to detail these

expenditure changes once we understand the true FTE budget needs. We do not expect any additional funds will be needed to complete this project.

Nothing to Report

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents

Nothing to Report

Significant changes in use or care of human subjects

Nothing to Report

Significant changes in use or care of vertebrate animals.

Nothing to Report

Significant changes in use of biohazards and/or select agents

Nothing to Report

6. PRODUCTS:

Publications, conference papers, and presentations

Dr. Timothy M. Kowalewski presented research related to skill evaluation methods to be used in this project at International Conference on Information Processing in Computer-Assisted Interventions:

"Predicting Surgical Skill from the First N Seconds" at the IPCAI conference, Barcelona, Spain June 2017

Journal publications.

UMN has submitted two papers related to development of the metrics and analysis processing applied in this study:

(*See appendix W81XWH-15-2-0030 Paper #1*)
"Predicting Surgical Skill from the First N Seconds of a Task Value over Task Time Using the Isogony Principle" Anna French, Thomas S. Lendvay M.D.,Robert M. Sweet M.D. , Timothy M. Kowalewski Ph.D., CARS International Conference on Information Processing in Computer-ssisted Interventions (IPCAI) 2017 [Published]

(*See appendix W81XWH-15-2-0030 Paper #2*)
"The Minimally Acceptable Classification Criterion for Surgical Skill: Intent Vectors and Separability of Raw Motion Data" Rodney L. Dockter, Thomas S. Lendvay M.D., Robert M. Sweet M.D., Timothy M. Kowalewski Ph.D., CARS International Conference on Information Processing in Computer-Assisted Interventions (IPCAI) 2017 [Published]

(*See appendix W81XWH-15-2-0030 Paper #3*)
"Laparoscopic Skill Classification Using the Two-Third Power Law and the Isogony Principle" Anna French, Timothy M. Kowalewski Ph.D, Journal of Medical Devices 2017 [Accepted]

Books or other non-periodical, one-time publications.

Nothing to report

Other publications, conference papers, and presentations.

Nothing to report

Website(s) or other Internet site(s)

Nothing to report

Technologies or techniques

Nothing to report

Inventions, patent applications, and/or licenses

Nothing to report

Other Products

Nothing to report

7. PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS

What individuals have worked on the project?

Name: Thomas Lendvay
No change

Name: Karen Edwards (No longer involved in this study)

Name: Anna French
No change

Name: Prof. Tim Kowalewski
No Change

Name: Sara Teller (No longer involved in this study)

Name: Lois Meryman (Assuming Sara Teller's' administrative role)
Project Role: Project Manager/Site Coordinator
Nearest person month worked: 2
Contribution to Project: Project management and subject management

***Example**:*

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?

Nothing to Report

What other organizations were involved as partners?

Nothing to Report

8. SPECIAL REPORTING REQUIREMENTS

COLLABORATIVE AWARDS: Nothing to Report

QUAD CHARTS: *See Appendix*

9. APPENDICES:

W81XWH-15-2-0030 Paper #1

Int J CARS manuscript No.
(will be inserted by the editor)

Predicting Surgical Skill from the First N Seconds of a Task

Value over Task Time Using the Isogony Principle

Anna French · Thomas S. Lendvay M.D. · Robert M. Sweet M.D. · Timothy M. Kowalewski Ph.D.

Received: date / Accepted: date

Abstract *Purpose:* Prior attempts at surgical skill evaluation have focused predominantly on diagnosis using task-specific maneuvers. These maneuvers required surgical expertise to identify and are observed over the course of a full task. The aim of this investigation is to propose features for automated skill evaluation that are relevant regardless of the surgical training task the tools perform. A secondary goal is to diagnose skill without requiring the complete time series of data from a given trial.

Methods: Features calculated from the isogony principle are used to train four common machine learning algorithms from dry-lab laparoscopic data gathered from three common training exercises. These models are used to predict the binary or ternary skill level of a surgeon. K-fold and leave-one-user-out cross-validation are used to assess the accuracy of the generated models.

Results: It is shown that the proposed scalar features can be trained to create 2-class and 3-class classification models that map to Fundamentals of Laparoscopic Surgery (FLS) skill level with median 85% and 63% accuracy in cross validation, respectively, for the targeted dataset. Also, it is shown that the 2-class models can discern class at 90% of best-case mean accuracy with only 8s of data from the start of the task.

Conclusion: Novice and expert skill levels of unobserved trials can be discerned using a state vector machine trained with parameters based on the isogony principle.

Anna French
Department of Mechanical Engineering University of Minnesota, Minneapolis, MN
E-mail: afrench@umn.edu

T. Lendvay
Department of Urology, Seattle Children's Hospital, Seattle, WA

R. Sweet
Department of Urology, University of Washington, Seattle, WA

T. Kowalewski
Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN

The accuracy of this classification comes within 90% of the classification accuracy from observing the full trial within 10 seconds of task initiation on average.

Keywords Surgical Skill Evaluation · Computer aided decision · Tracking systems

1 Introduction

Surgical technical skill directly impacts patient health outcomes, as shown in [2]. An accurate automated surgical skill evaluation system would consequently be an important tool in reducing a surgical patient's injury risk. A system able to deliver evaluations immediately following a training module would also prove beneficial to surgeons in training, since formative feedback is a lauded goal in surgical training [12]. Knowing when one makes an error is invaluable information that molds good behavior. It is a core educational principle that spans disciplines [13]. Current methods introduce time barriers to feedback, either by requiring humans to deliver ratings or by using scoring models which require the compilation of scored trial data each time a new training technique is developed. To accomplish timely feedback, a scoring method must be developed that depends neither on human intervention nor on prior probability distributions for features specific to a particular task.

Past attempts at developing an automated skill evaluation system have focused on diagnosing skill using task-specific performance measurements. In [10], a method using linear discriminant analysis and tool motion features achieved accuracy in the 90% range, however the assessment method was tailored to four-throw suturing tasks and defining task features (surgemes) required a surgeon's expertise. Investigation from [1] reports similar classification success but also used features very specific to septoplasty, and classified by segmenting and analyzing the stroking motions of the cottle. The investigation in [8] showed the crowd is capable of discerning surgical skill concordant with the current gold standard (an expert panel) and can generate a skill evaluation weeks faster than the expert panel. However, these scores still depend on human intervention and also introduce several hours of lag time between task completion and score delivery to the trainee. A method that does not require task completion to diagnose skill and can model skill regardless of which task the surgeon performs would be preferred.

The Fundamentals of Laparoscopic Surgery is a high-stakes certification exam of cognitive and technical laparoscopic skills [4, 5, 11]. It is now often required for graduation and or board certification among laparoscopic curricula. It has been extensively validated and even shown to correlate with patient outcomes [14]. However, there are some limitations. The score is based on task time and penalty counts (e.g. dropping an item, cutting outside a boundary, loose knots). [7] found that the weights used in computing FLS score greatly emphasize task time, rendering the penalties virtually irrelevant. This suggests that FLS score provides little or no practical value over task time. However the value of FLS scoring (or, implicitly, FLS task time) in its link to patient outcomes remains undisputed and therefore valuable as a measure of surgical technical skill. Our observations in operating rooms, surgical simulator sessions, and among trainees reveal that a subject's approximate level of technical skill is often evident very

quickly–within tens of seconds–when watching video of their tool motions during a procedure. This suggests that skill evaluation can be correctly approximated with a fraction of the time it takes to do a procedure–be it an FLS task or a surgical procedure. This would alleviate the need for trainees and proctors to wait until the end of a task to receive a skill evaluation. However, it is unclear either what motion features to identify or how much time is needed before obtaining adequate confidence in such an assessment, i.e., how many seconds are required to predict an FLS score?

1.1 Isogony Principle

The isogony principle may provide some value to tool motion-based skill evaluation. In [9], subjects were recorded drawing shapes of various curvature with the goal of relating curvature of the drawn shape to speed of the pen tip. A relation between these two parameters was determined using the isogony principle as:

$$v(t) = \gamma k(t)^{1/3}$$

where $v(t)$ is the instantaneous velocity of the tip of the pen, $k(t)$ is the local curvature that the tip of the pencil traces, and γ is the velocity gain factor parameter relating $v(t)$ and $k(t)$. In [9], it was asserted that $v(t)$ can be predicted from $k(t)$ based on a constant value of γ for a given segment of motion.

This investigation extends the velocity gain factor relationship to 3D tool motion, using the velocity and curvature from the 3D space. For the purpose of this study, we do not assume constant values of γ, and instead choose to observe the behavior of the γ parameter:

$$\gamma(t) = \frac{v(t)}{k(t)^{1/3}}$$

For the $k(t)$ parameter, the radius of curvature was used:

$$k(t) = \frac{(1 + v(t)^2)^{3/2}}{a(t)}$$

1.2 Hypotheses and Objectives

Based on the property from [9], several hypotheses were drawn. First, it was hypothesized that the variability of the γ parameter between novice surgeons will be small. This was drawn from the idea that novice surgeons will adhere more to their "natural" hand motion pattern, while the more practiced motions of experienced surgeons will vary from this natural motion pattern.

Second, it was hypothesized that scalar parameters such as the mean (μ) and standard deviation (σ) of a trial's γ for each hand can be used as features to train machine learning algorithms and coarsely predict the Fundamentals of Laparoscopic Surgery (FLS) score of unobserved trials.

Third, it was hypothesized that the full duration of the task is not required to evaluate skill since $\gamma(t)$ is easily observable at any point in the task. Accordingly,

investigation was made into the minimum number of seconds of data from a trial required to discern the subject's FLS class with an acceptable level of accuracy. This ability would provide significant value over task time related features, which require probability distributions based on the results of previous users, and would not be agile to changes in training. A secondary hypothesis is that prediction accuracy will increase as more time is included, but gains will taper off.

Results from both task-specific and task-blind models were generated and are reported here. Task-specific models are models trained using only samples from a specific task, and are included as a basis for comparison. Task-blind models are models where data across all tasks were included in training, and were used to predict the skill of any task. Both leave-trial-out cross validation and leave-one-user-out (sometimes referred to as leave-surgeon-out, and abbreviated here as LOUO) validation methods were used to test the accuracy of the binary and ternary classification algorithms developed.

There are three key contributions offered by this paper. First, we introduce isogony as a potentially useful feature in surgical skill evaluation. Second, we introduce task invariance as a desired attribute of skill evaluation. Third, we introduce the notion of estimating skill normally evaluated over the course of a full task from a partial task observation, i.e. predicting final scores from N seconds.

2 Methods

2.1 Dataset

This investigation used the dataset established in the Electronic Data Generation and Evaluation (EDGE) study described by [7]. This study gathered video, tool motion and demographic data on 98 different surgeons performing typical FLS tasks. From this data set, 108 peg transfer, 63 suturing and 124 circle cutting tasks were used for this study.

Each instance where data was recorded while a subject was performing a particular FLS task will be referred to as a "trial." Within the dataset, each trial is comprised of a 30Hz fixed camera-position video recording of the laparoscopic tools interacting with the training field, numeric data documenting the position, orientation and grasp force of the tool tips corresponding to each frame in the video, an FLS score ranking the subject's skill level based on their performance in each trial, and demographic information relating relevant information about the subject such as their dominant hand and experience level.

The FLS score alone was used to establish skill groups within each of the three tasks. This resulted in the FLS expert class (any trial with FLS scores above a threshold of OSATS scores from identity-blind review by two faculty surgeons for each task; see [6]); the FLS novices (trials from the bottom 15th-percentile of FLS scores within each task); and FLS Intermediates (trials from the 15th-percentile range about a midpoint between the lowest FLS Expert score and highest FLS Novice score, for each task). The video portion of the data was not used. This choice of criteria gave us complete trials from 67 FLS Novices, 71 FLS Intermediates, and 157 FLS Experts.

2.2 Analysis Methods and Algorithms

The mean and standard deviation of the γ parameter of each trial for the dominant (d) and non-dominant (nd) hands are the features selected for evaluation. These are referred to as $\sigma(\gamma_d), \mu(\gamma_d), \sigma(\gamma_{nd})$ and $\mu(\gamma_{nd})$. These four features were calculated for each trial, and were used along with their FLS class to train several different machine learning algorithms to classify skill level. The accuracies these trained algorithms obtained were used as evidence of feature strength. The algorithms used for testing were logistic regression (LR), support vector machine (SVM), linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA).

The accuracy of each trained model was evaluated using k-fold cross validation (with k=10) and leave-one-user-out validation. For the k-fold, each fold was assigned an equal number of trials from each class. The k-fold cross validation was performed N=10 different times, where a new set of k-folds was selected and evaluated for each iteration of N, which resulted in kN different models trained and evaluated for each machine learning algorithm. Note that these are folds created using each individual trial, hence it is partitioning in a leave-trial-out manner.

Accuracy was also evaluated using leave-one-user-out (LOUO) for all Q surgeons. Each surgeon has r different trials in the database, where r may differ for each surgeon. In this method, each surgeon takes a turn as the test set while the other $Q-1$ are used for the training set to generate the models. The accuracy is reported by evaluating the classification results of each of the r trials for each of the Q surgeons. Feature strength and model accuracy were assessed separately for both 2-class classification (discriminating between novice and expert) and 3-class classification (discriminating between novice, intermediate, and expert).

Models were generated in both task-specific and task blind manners. Task-specific models were trained using only trials where a specific task was performed, and their accuracy was tested using only trials from that specific task. Task-blind models were trained using all trials regardless of task and were used to create predictions of any trials regardless of class. The accuracy of task-blind model predictions for each specific class was also analyzed, where the model was trained task-blind but the testing set was partitioned to analyze how well the task-blind model can predict the skill for each specific task.

The minimum period of time required for acceptable prediction accuracy was evaluated by taking successively longer series of time from the beginning of each trial to time t and calculating $\mu({}^t\gamma_d)$, $\sigma({}^t\gamma_d)$, $\mu({}^t\gamma_{nd})$, and $\sigma({}^t\gamma_{nd})$ based on those different time periods. The feature $\mu({}^t\gamma_d)$ is the feature $\mu(\gamma_d)$ calculated from the γ values from the start of the training exercise until time t, where $n(t)$ represents the number of time-steps included in the range [0,t] (data was recorded at 30Hz, so $n = 30t$):

$$\mu({}^t\gamma_d) = \frac{1}{n(t)} \sum_{i=0}^{n(t)} \frac{v_d(i)}{k_d(i)^{1/3}} \qquad \sigma({}^t\gamma_d) = \sqrt{\frac{1}{n(t)} \sum_{i=0}^{n(t)} \left(\frac{v_d(i)}{k_d(i)^{1/3}} - \mu({}^t\gamma_d) \right)}$$

The $\mu({}^t\gamma_{nd})$ and $\sigma({}^t\gamma_{nd})$ are calculated by the same method, but with the non-dominant hand measurements. The above features were generated for each integer-valued time period second within $t = (1, 30)s$, where 30s was chosen since all task times in the data set were greater than this amount. This created 30 different

Table 1: Median and standard deviation of the $\sigma(\gamma_d)$ and $\sigma(\gamma_n d)$ over all subjects.

		$\sigma(\gamma_d)$ med(standard deviation)	$\sigma(\gamma_{nd})$ med(standard deviation)
PegTx	Nov	13.91 (4.22)	11.52 (5.83)
	Exp	22.86 (7.03)	20.98 (8.71)
Cutting	Nov	9.24 (5.83)	9.72 (3.88)
	Exp	18.37 (5.97)	18.83 (10.61)
Suturing	Nov	9.09 (2.31)	9.45 (2.76)
	Exp	14.85 (5.83)	17.66 (7.90)
All Tasks	Nov	10.19 (5.14)	9.93 (4.23)
	Exp	18.80 (7.04)	19.45 (9.57)

groups of $\mu({}^t\gamma_d)$, $\sigma({}^t\gamma_d)$, $\mu({}^t\gamma_{nd})$, and $\sigma({}^t\gamma_{nd})$ specific to the period of time they were calculated from. Each of these 30 groups were then passed through the same machine learning algorithms and validation process as for $\sigma(\gamma_d), \mu(\gamma_d), \sigma(\gamma_{nd})$ and $\mu(\gamma_{nd})$, yielding a mean accuracy μ_t for each group. Trends for the value of μ_t for $t = (1, 30)s$ for each different machine learning algorithm were then plotted, and are displayed in the results section. The minimum t required to get within 90% of the observed settling accuracy is reported in Table 2.

In addition, these methods have been validated against other validated methods for skill classification, such as those in [3]. The validation methods trains three different models using either tool path lengths (PL), economy of motion (EOM) and motion smoothness (MS) as features using SVM.

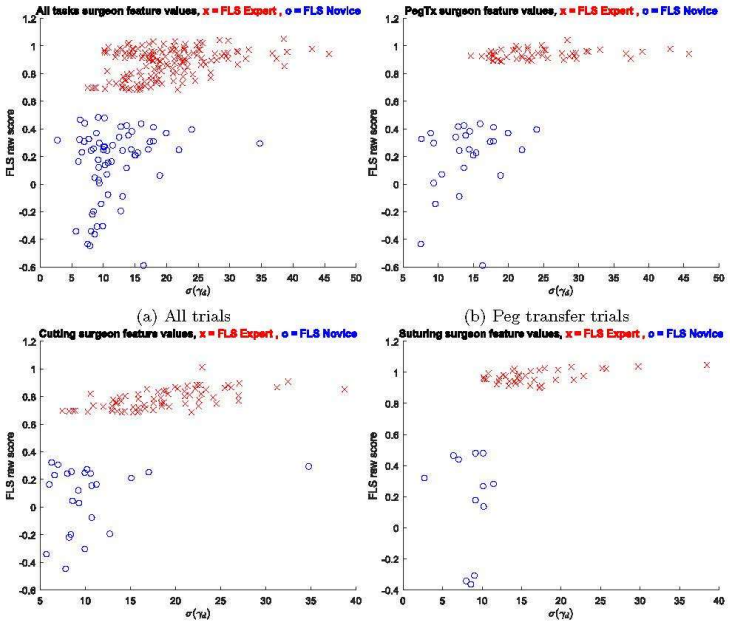
3 Experimental Results

3.1 Two Class Classification: FLS Expert vs. FLS Novice

Figures 1a-1d show the distribution of FLS scores plotted against the $\sigma(\gamma_d)$ feature of each trial for a given subject, where the marker type and color specifies expertise. Recall that $\sigma(\gamma_d)$ and $\sigma(\gamma_{nd})$ are features representing the intra-subject standard deviations (the subject's standard deviation for motion during a given trial). Table 1 details the median and range of the $\sigma(\gamma_d)$ and $\sigma(\gamma_{nd})$ features for each subject. It is observable from here that the inter-subject medians for novices are much lower and have much smaller inter-subject standard deviation.

Figures 2a-2c show six example plots of the γ_D parameter calculated for each time step in the first 20 seconds of a trial.

Figure 3a shows the statistics reporting the mean accuracy of each model trained to classify between FLS novice and FLS expert trials. Five different types of models were generated and tested at each round. The left four box plots represent the model prediction accuracy based on training using a single feature. The rightmost column used all four features to train the model. The mean accuracy of each model trained in each k-cross validation and each N-iteration was recorded and used to generate the box plots. Thus the statistics displayed are values calculated over the kN models generated in the cross validations. These models were trained task-blind, meaning all trials regardless of task type were used to train the model. They were also tested task-blind, meaning the results reported here are the accuracy over all trials regardless of task.



(a) All trials (b) Peg transfer trials

(c) Cutting trials (d) Suturing trials

Fig. 1: FLS score vs $\sigma(\gamma_d)$ feature calculated over the full duration of each trial (dominant hand). Each point represents the feature value for one trial. Raw data displayed for all trials and each trial individually.

For all provided box plots, the 25th and 75th percentiles are the lower and upper box boundaries while the median is the central line. The whiskers extend to the most extreme non-outlier points, and the + are considered outliers. The models trained using single features from each hand have agreement not far from the model trained on all features, which shows a median agreement with desired skill class of 85%.

Figure 4a shows the box plots representing model prediction accuracy using LOUO. These models are trained using all four features and the labeled machinery type. This figure shows a median model prediction accuracy of between 80% and 100% depending on model type. Note that LOUO tests the accuracy of each user individually, i.e. each user has performed n trials and a prediction accuracy is assessed for each individual user based on the percentage of those n trials that were correctly classified. The LOUO box plot displays the crowd tendencies of the percent accuracies of each user.

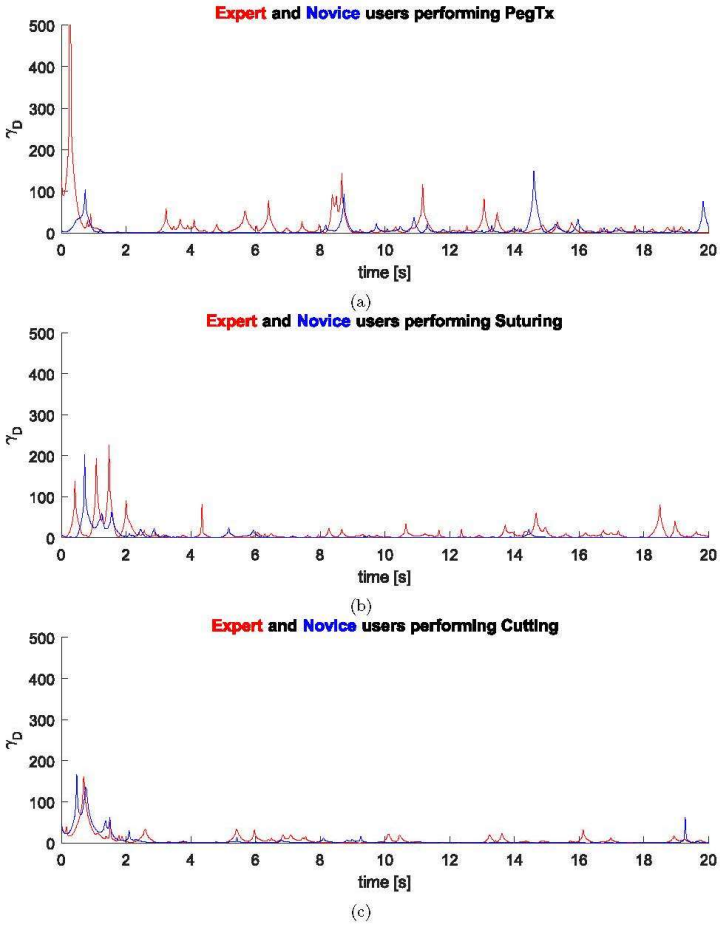
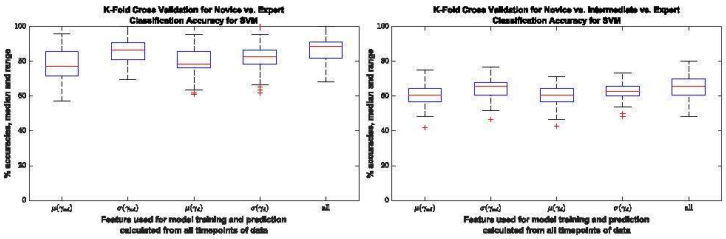


Fig. 2: Example of $\sigma(\gamma_d)$ activity over the first 20 seconds of each trial

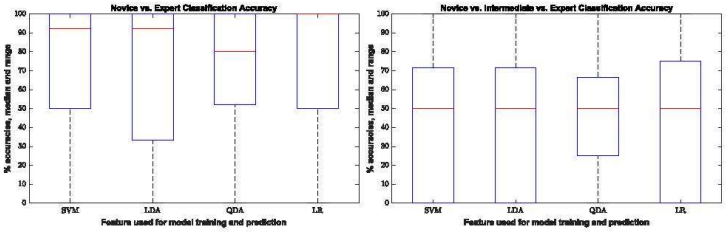
3.2 Three Class Classification: FLS Expert, FLS Intermediate, FLS Novice

Figure 3b uses the same k-fold method as Figure 3a, however it predicts over all three classes (novice, intermediate, expert) rather than just between novice and expert. It shows a median agreement with desired skill class of 62% when trained



(a) 2-class classification (b) 3-class classification

Fig. 3: K-fold leave-trial-out SVM classification task-blind training and task-blind testing, comparing strength of each feature.



(a) 2-class classification (b) 3-class classification

Fig. 4: LOUO classification task-blind training and task-blind testing using $\mu({}^{t}\gamma_d)$, $\sigma({}^{t}\gamma_d)$, $\mu({}^{t}\gamma_{nd})$, and $\sigma({}^{t}\gamma_{nd})$ features combined to train each model, comparing strength of machinery.

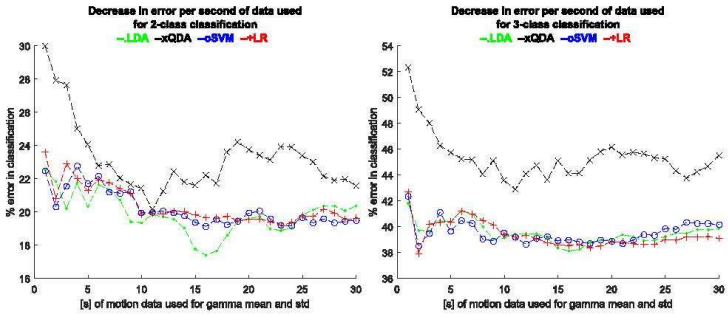
using all features. Figure 4b uses the same LOUO method as Figure 4a, and it also predicts over all three classes. It shows a median agreement of 50%.

3.3 Minimum Time to Classification

Figure 5a was generated from SVM, QDA, LR and LDA models trained using all four features and shows the overall model error rate of as t is increased from 0s to 30s, which increases the number of data points in a trial used to calculate $\mu({}^{t}\gamma_d)$, $\sigma({}^{t}\gamma_d)$, $\mu({}^{t}\gamma_{nd})$, and $\sigma({}^{t}\gamma_{nd})$. Integer values of t from 1 to 30 were used. Figure 5b uses similar methods, but is classifying between novice, intermediate and expert surgeons.

3.4 Validation with Similar Methods

Figure 6a and 6b compare the performance for 2-class classification of the γ parameters against previously validated aggregate task metrics as described in [3].



(a) 2-class classification (b) 3-class classification

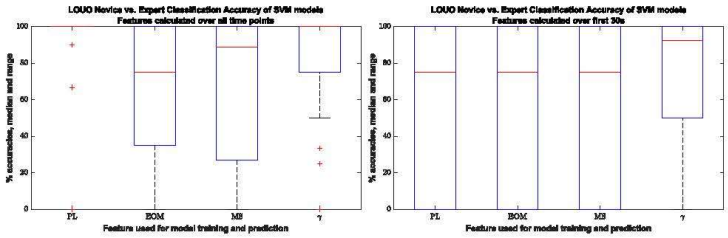
Fig. 5: Error rate per number of seconds used to calculate $\mu({}^t\gamma_d)$, $\sigma({}^t\gamma_d)$, $\mu({}^t\gamma_{nd})$, and $\sigma({}^t\gamma_{nd})$.

Table 2: 2-class classification accuracy results per algorithm and per task. Models generated for this figure were trained task-blind, results at left reports how well the model classifies each task

	Task Specific accuracy (μ) over all kN model accuracies in cross validation, mean(std dev)			Task Blind mean accuracy over all kN model accuracies in cross validation	
	PegTx $\mu(\sigma)$	Cutting $\mu(\sigma)$	Suturing $\mu(\sigma)$	Best % Accuracy Mean/Median	Min time to 90% of best accuracy
LDA	80.1 (.13)	86.1 (.11)	83.0 (.2)	83.7/85.7	2s
QDA	87.6 (.12)	86.9 (.11)	66.0 (.24)	82.5/82.6	4s
SVM	85.2 (.12)	90.2 (.10)	82.6 (.20)	87.2/88.7	8s
LR	84.6 (.12)	90.2 (.10)	81.5 (.21)	86.6/86.4	7s

Table 3: 3-class classification accuracy results per algorithm

	Task Specific accuracy (μ) over all kN model accuracies in cross validation, mean(std dev)			Task Blind mean accuracy over all kN model accuracies in cross validation	
	PegTx $\mu(\sigma)$	Cutting $\mu(\sigma)$	Suturing $\mu(\sigma)$	Best % Accuracy Mean/Median	Min time to 90% of best accuracy
LDA	57.3 (.13)	63.0 (.14)	66.0 (.16)	61.6/61.7	2s
QDA	66.6 (.12)	59.2 (.14)	44.3 (.16)	58.9/58.6	4s
SVM	62.2 (.13)	67.0 (.13)	65.5 (.14)	65.1/65.5	3s
LR	60.8 (.12)	68.3 (.13)	65.7 (.15)	65.1/64.3	3s



(a) Features calculated at end of task time (b) Features calculaed at 30s point in task time

Fig. 6: LOUO validation for a 2-class classification using SVM for task-blind training and task-blind testing, comparing strength of each feature set.

PL_d and PL_{nd} are the path length variables calculated for the dominant and non-dominant hand. The PL boxes represents the accuracy of an SVM trained using PL_d and PL_{nd} together as features, with accuracy measured using k-fold for N iterations. The same applies for EOM and MS boxes. The γ box represents the accuracy of an SVM trained using $\mu(\gamma_d)$, $\sigma(\gamma_d)$, $\mu(\gamma_{nd})$, and $\sigma(\gamma_{nd})$.

4 Conclusion

The results give support to our three initial hypotheses. The first hypothesis is supported by Figures 1a - 1d and Table 1. The feature $\sigma(\gamma_d)$ is taken to represent the intra-subject variability in the γ_d parameter for a hand. A low intra-subject γ_d (i.e. a small $\sigma(\gamma_d)$), may imply a given subject is nearly following the motion law outlined in [9], where the tool tip is assumed to maintain a constant γ_d. Broadening the scope to how these skill levels behave at the group level, from Table 1 it is observable that novice subjects have inter-subject median and standard deviation values for $\sigma(\gamma_d)$ and $\sigma(\gamma_{nd})$ that are comparatively lower than experts. This supports the first hypothesis. A comparatively low inter-subject standard deviation for the $\sigma(\gamma_d)$ feature for the novices may imply a behavioral pattern between subjects. Meanwhile, a comparatively low inter-subject median for the $\sigma(\gamma_d)$ feature for the novices may imply that, as a group, novices stay closer to the "natural" motion law. This could suggest that experienced surgeons mature out of this adherence to the motion pattern with practice for laparoscopic tools.

Second, the scalar parameters $\sigma(\gamma_d)$, $\mu(\gamma_d)$,$\sigma(\gamma_n d)$, and $\mu(\gamma_n d)$ were able to train LDA, QDA, SVM and LR models to predict the class of partitioned data with mean cross validation accuracy in the 85% region for binary classification and in the 60% region for ternary classification. Prediction accuracy using LOUO yielded median accuracy of up to 100% for binary classification and 50% median accuracy for ternary classification using Logistic Regression. It should be noted that there is a large variation in classification accuracy across the different users for the LOUO box plots. Chiefly, for a 2-class classification the 25th percentile is as low as 33% for certain users while a 3-class classification gets all the way to 0% for certain users. The outliers also reach 0% for 2-class classification. Further

investigation will have to be made into this behavior. It is unclear at this point whether this exposes a limitation in the chosen features or whether our data still has insufficient N to capture human variability in surgery.

Third, Figure 5a, Figure 5b, Table 2 and Table 3 show that for all four tested algorithms, the time required to get within 90% of the best observed accuracy is less than the full task duration. Note there is some oscillation in several of the curves in Figure 5a. It is not obvious what this signifies, e.g. data may be truncated through incomplete maneuvers or γ may be only significant at sustained speeds to rise above the noise floor. Exploring this will require a dataset with motion segments continuously labeled by skill level.

In addition, this method was compared against previously validated methods. For the EDGE dataset, Figures 6a and 6b show that γ parameters outperform economy of motion and motion smoothness, but fall short of path length. When measured short of task completion (at 30s), Figure 6b shows that γ parameters outperform the validation features. This is expected, since the validation metrics are heavily influenced by task time, a characteristic γ features are free of.

Determining the FLS class in the first seconds of a task for this dry-lab simulation data is a significant outcome. This implies that a trainee and proctor can potentially take less time for FLS certification. Also, traditional human-required tasks in FLS penalty scoring that were resource intensive – such as counting object drops or measuring cut accuracies – may not always be required. The positive classification results that were done in task-blind settings also suggest that the isogony measure may be capturing some of the aspects of skill evident in human motion that may be obvious to expert reviewer but difficult to articulate– aspects that may allow them to infer skill from only a few seconds of a video.

Prior art has typically not investigated task-blind skill classification methods. Our results of median 85% accuracy for novice-expert classification within the first 30s of a task rival or outperform existing, often more complex approaches. The fact that a task-blind model can be generated using γ with the demonstrated accuracy suggests that γ provides insight into some task invariant attributes of skill level.

There were several limitations in this study that should be addressed in future work. This study used only dry-lab laparoscopic simulation tasks, which do not necessarily mimic real surgical maneuvers. This limits our results and conclusions to only this simulated manual laproscopy context. The skill groups used here are defined based on FLS score only and are thresholded based on the subjects available in the dataset. Data defining skill based on the surgical panel and crowd sourced skill determinations will be used in its place in the future. It was assumed that meaningful motion was occurring in the data used this study, and selecting only the first 30s of task execution was used as a surrogate for capturing meaningful motion. It is possible that some trials may include subjects keeping their tools immobile while planning their maneuvers at the start of the task, so this must be filtered out in future work. Additional datasets could also be generated by sub-sampling randomized time intervals from existing tests. This would also help investigate the question of whether the quicker diagnoses (within 2 seconds) are due to the fact that expert surgeons get to work more quickly and confidently early in the task than novices.

We do not claim that our approach, as given, is immediately useful to surgical trainees. However, it is a necessary step towards achieving formative feedback. Namely, if a skill measuring feature only correlates with task time (e.g. FLS score

is almost identical to task time [7]), it would have little or no value for formative feedback (or even as a summary metric itself). We show that isogony provides some accuracy in measuring skill even within the first N seconds, this suggests that it has some utility over task time. However, this is a necessary but not sufficient step for formative feedback. For example, a mapping of isogony features to easy-to-understand continuous motion quality scores on, say, a percentage scale could be more useful.

We conclude that predicting final FLS score from roughly the first 10 seconds of a trial is potentially feasible and that isogony provides some useful task-blind skill-classification information above simple task-time or FLS score.

Compliance With Ethical Standards

Conflicts of Interest The authors declare that they have no conflict of interest.
Ethical Standard All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
Funding This work was supported, in part, by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-15-2-0030. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the Department of Defense.
Informed Consent Informed consent was obtained from all individual participants included in the EDGE study.

References

1. Ahmidi N, Poddar P, Jones JD, Vedula SS, Ishii L, Hager GD, Ishii M (2015) Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. International journal of computer assisted radiology and surgery 10(6):981–991
2. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369(15):1434–1442
3. Chmarra M, Kolkman W, Jansen F, Grimbergen C, Dankelman J (2007) The influence of experience and camera holding on laparoscopic instrument movements measured with the trendo tracking system. Surgical endoscopy 21(11):2069–2075
4. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL (1998) Development of a Model for Training and Evaluation of Laparoscopic Skills. The American Journal of Surgery 175(6):482–487
5. Fried GM (2008) FLS Assessment of Competency Using Simulated Laparoscopic Tasks. Journal of Gastrointestinal Surgery 12(2):210–212, DOI 10.1007/s11605-007-0355-0, URL http://dx.doi.org/10.1007/s11605-007-0355-0
6. Kowalewski TM (2012) Real-time quantitative assessment of surgical skill. PhD thesis, University of Washington

7. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet R, Wright A, Hannaford B, Sinanan MN (2014) Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology. Journal of Surgical Research 192(2):329–338
8. Kowalewski TM, Comstock B, Sweet R, Schaffhausen C, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Lee B, Schwartz BF, McDougall E, Lendvay TS (2016) Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. The Journal of urology 195(6):1859–1865
9. Lacquaniti F, Terzuolo C, Viviani P (1983) The law relating the kinematic and figural aspects of drawing movements. Acta psychologica 54(1-3):115–130
10. Lin HC, Shafran I, Yuh D, Hager GD (2006) Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. Computer Aided Surgery 11(5):220–230
11. Peters JH, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, Hoffman K (2004) Development and Validation of a Comprehensive Program of Education and Assessment of the Basic Fundamentals of Laparoscopic Surgery. Surgery‘ 135:21–27
12. Rogers GM, Oetting TA, Lee AG, Grignon C, Greenlee E, Johnson AT, Beaver HA, Carter K (2009) Impact of a structured surgical curriculum on ophthalmic resident cataract surgery complication rates. Journal of Cataract & Refractive Surgery 35(11):1956–1960
13. Shute VJ (2008) Focus on formative feedback. Review of educational research 78(1):153–189
14. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM (2010) Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room – a randomized controlled trial. The American journal of surgery 199(1):115–120

Int J CARS
DOI 10.1007/s11548-017-1610-9

ORIGINAL ARTICLE

The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data

Rodney L. Dockter[1] · **Thomas S. Lendvay**[2] · **Robert M. Sweet**[3] · **Timothy M. Kowalewski**[1]

Received: 27 January 2017 / Accepted: 8 May 2017

Abstract
Purpose Minimally invasive surgery requires objective methods for skill evaluation and training. This work presents the minimally acceptable classification (MAC) criterion for computational surgery: Given an obvious novice and an obvious expert, a surgical skill evaluation classifier must yield 100% accuracy. We propose that a rigorous motion analysis algorithm must meet this minimal benchmark in order to justify its cost and use.
Methods We use this benchmark to investigate two concepts: First, how separable is raw, multidimensional dry laboratory laparoscopic motion data between obvious novices and obvious experts? We utilized information theoretic techniques to analytically address this. Second, we examined the use of intent vectors to classify surgical skill using three FLS tasks.
Results We found that raw motion data alone are not sufficient to classify skill level; however, the intent vector approach is successful in classifying surgical skill level for certain tasks according to the MAC criterion. For a pattern cutting task, this approach yields 100% accuracy in leave-one-user-out cross-validation.
Conclusion Compared to prior art, the intent vector approach provides a generalized method to assess laparoscopic surgical skill using basic motion segments and passes the MAC criterion for some but not all FLS tasks.

✉ Rodney L. Dockter
dockt036@umn.edu

[1] Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN, USA

[2] Department of Urology, Seattle Children's Hospital, Seattle, WA, USA

[3] Department of Urology, University of Washington, Seattle, WA, USA

Published online: 18 May 2017

Keywords Surgical skill evaluation · Surgical training · Surgical motion · Laparoscopic surgery

Introduction

The fundamentals of laparoscopic surgery (FLS) were developed to evaluate and credential laparoscopic surgeons. The FLS scoring criteria are based primarily on task time and number of task errors as determined by a qualified proctor. While FLS has been shown to discriminate between expert and novice subjects [18], these measures have the potential to miss key information and overemphasize task time [13]. The challenges related to laparoscopic surgery motivate the development of objective, automated, and accurate surgical skill evaluation techniques.

Prior work on surgical skill evaluation has been widespread. One approach has utilized aggregate task measures such as task time and path length [5,6]. In [16], task level metrics were used to estimate pairwise maneuver preferences with 80% accuracy. In [9], robotic arm vibrations and interaction forces were used within a composite skill rating; however, statistical analysis showed that completion time provided the primary contribution. Another method has been to decompose surgical tasks into specific gestures or 'surgemes' [15]. Using these surgemes, models for skill can be trained using a variety of machine learning approaches. Hidden Markov models (HMMs) have been used extensively to model surgical skill level. An HMM model for various surgemes was used to classify a sequence as a particular skill level [17]. This resulted in 100% classification accuracy for leave-one-super-trial-out (LOSO) cross-validation but required manually segmented surgemes and did not report leave-one-user-out (LOUO) validation results. The results of [19] had high classification rates for LOSO cross-

validation, but these results fell precipitously under LOUO validation suggesting overfitting. Another method utilizes descriptive curve coding (DCC) in which the principal direction changes within a trajectory are encoded as a string of integers [1]. With this approach, encoded common strings or motifs were used to model skill level. This method results in 98% accuracy for LOSO validation but around 90% for LOUO. Task-specific motion models have been proposed for procedures such as septoplasty [2]. This approach involves stroke-based features designed to assess the consistency and efficiency with which a surgeon removes skin from underlying cartilage. This approach gave a LOSO classification accuracy of 90%, but its applicability to other procedures is not yet clear. The ribbon area measure treats the surgical tool wrist as a brush and measures the accumulated surface area of the trajectory as a surrogate for dexterity [11]. This approach resulted in an 80% binary classification accuracy. Both the stroke-based features and ribbon area approaches are conceptually similar to the work presented here; however, we attempt to use these concepts in a manner more generalizable across tasks and that results in a higher classification accuracy. The gap in prior art has been a fully automated algorithm which provides 100% classification between obvious expert and novice surgeons using LOUO cross-validation.

Prior art has revealed a secondary problem: Data set categories are unreliably labeled relative to true skill level. These categories are typically defined by subject demographics such as caseload, academic rank, or experience level. Yet even an expert surgeon can exhibit skill decay and demonstrate a variance in skill level within a given context. True experts or technical masters can sometimes (e.g., for a given grasp or motion within an entire procedure) exhibit novice-like motions. Kowalewski et al. [14] showed that expert categories based on these demographics are unsuitable for validation studies as they often result in recorded trials from perceived experts that can exhibit poor technical skill. Overall this can confound supervised classifiers that assume a clean ground truth for correct analysis. The current gold standard for skill assessment is blinded review of surgical videos by panels of expert surgeons using structured survey tools such as the objective structured assessment of technical skill (OSATS) [7]. Birkmeyer et al. [3] showed that using similar evaluation methods technical skill can be linked directly with patient outcomes. To this end, Kowalewski et al. [13] defined a ground truth expert trial (a single recording by a given individual) as one that is deemed an expert by a consensus of three validated methods: demographically-derived expertise, FLS score, and OSATS-like video review.

We herein introduce the minimally acceptable classification (MAC) criterion for computational skill evaluation: Given an obvious novice and an obvious expert, the classification accuracy must be 100%. Some misclassification may be acceptable between other skill levels, e.g., experts versus Master or Intermediate versus expert, but not an *obvious* novice versus *obvious* expert. Here we define obvious novices as subjects who should never be allowed to operate (always disqualified) and obvious experts as subjects who should never be disqualified from operating. Surgery requires this stipulation given that patently unqualified surgeons endanger lives. Often, such a large difference is very evident via task time or a casual viewer watching a video [4]. Therefore, a rigorous motion analysis algorithm should meet this minimal performance benchmark in order to justify cost and use. While this is not a sufficient criteria, it does provide a *minimal* necessary criterion to use as a baseline in this field. Our approach in this study was twofold. First, we asked 'how valuable is raw tool motion data alone in classifying skill given the MAC criterion?' Second, we present the 'intent vectors' feature and classification scheme applied to laparoscopic tool motion. We tested the hypothesis that intent vectors successfully classify skill according to the MAC for specific tasks.

Methods

In this section, we present the data set utilized in this study, the separability analysis used to assess raw surgical motion data, and the intent vectors derivation. The lack of separability in the raw data motivates the intent vectors.

Data set

This study utilized a previously recorded data set [13] where the electronic data generation for evaluation (EDGE) platform (Simulab Corp., Seattle, WA, USA) was used to collect task video data and tool motion data from participants including surgical faculty, residents, and fellows. Participants in the study performed a subset of the FLS tasks; peg transfer, pattern cutting, and intracorporeal suturing. Each subject was asked to complete, at minimum, three iterations of the peg transfer task, two iterations of the pattern cutting task, and two iterations of the suturing task. The subject pool consisted of 98 total subjects from a variety of specialties including General Surgery, Urology, and Gynecology spanning three teaching hospitals. Two FLS-certified graders manually recorded task errors, and task completion time was automatically recorded. Task errors and completion time were then used to compute an overall FLS score for each iteration.

From this data set, we have chosen the ground truth expert group (determined by a combination of caseload, FLS score, and p-OSATS score) for our 'obvious expert' category and the FLS novice group (determined by the bottom 15th percentile of FLS scores for trials in each task) for our 'obvious novice' category. Individuals with such low scores would fail FLS and thus not be allowed to operate. The complete data

Table 1 FLS trials by task and skill level

Skill level	Peg transfer	Pattern cutting	Suturing
'Obvious novice'	29	25	13
'Obvious expert'	6	10	8

set contains 447 recorded trials across three tasks [13]. We selected only 91 of the original recorded trials to represent the extremes of 'obvious experts' and 'obvious novices.' Each trial was performed by a different subject (Table 1).

Each task was recorded with time synchronized video and tool motion data. This provided time-stamped Cartesian positions (x, y, z in cm) along with tool roll and grasper jaw angle (θ, degrees) at 30 Hz. This allowed subsequent computation of motion derivatives such as velocity and acceleration. In post-processing, surgical tool motion was segmented into distinct motions within each task based on information from the tool grasper at the distal end. A segment was considered to begin when the grasper was opened ($\theta > 3°$) and the force within the grasper jaws fells below a threshold ($F_g < 4N$). The segment was then considered complete when the jaws were closed ($\theta < 3°$) and the force applied within the grasper jaws rose above a threshold ($F_g > 4N$) for 200 ms [13]. Each tool is segmented separately, allowing for overlapping segments between each instrument (hand). The mean number of segments per trial and the mean segment duration are given in Table 2.

Functionally this segmentation scheme results in segments where a tool is moved in a trajectory toward an object, and then the jaws are closed around the object to secure it, thus ending the segment. Our segments focused only on tool motion where the surgeon is reaching toward an object (e.g., before grasping or cutting), a motion which is prevalent in nearly all surgical tasks. The goal of this segmentation scheme was to be generalizable to all surgical tasks as compared to task-specific surgical gestures. We expected that some spurious false positives may occur within segmentation and assumed that these false segments occur equally across skill groups.

Value of 'raw motion data' for classification

To explore the separability of dexterous skill levels given raw motion data from EDGE, we refined and utilized information theoretic techniques, starting with the RELIEFF algorithm [12]. This is used in binary classification to rank features based on their ability to separate the data effectively. For each point, we find the K-nearest neighbors belonging to the true class (hit) and the opposite class (miss). Using these nearest neighbors, a mean distance to both the hit neighbors (D_{hit}) and the miss neighbors (D_{miss}) is computed. The weights for a particular feature (W_f) are updated according to the difference between mean hit distance and mean miss distance (computed using that particular features data) (Eq. 1).

$$W_f = \sum_{i=1}^{N} \left(D_{\text{hit}_i} - D_{\text{miss}_i} \right). \tag{1}$$

Once weights for each feature have been computed, the features are sorted based on weight. Features with the highest weights are considered the most relevant features for classification. RELIEFF and its variants are limited to considering each feature separately and do not consider combinations of features simultaneously.

An obvious extension of the RELIEFF approach for multiple features (a variant termed RELIEF-RBF) utilizes radial basis functions (RBF) to estimate the probability density function given within class (hit) and between class (miss) data across any combination of n dimensions. As compared to the standard RELIEFF approach, all data from all dimensions contribute to the overall probability of that data point instead of only considering nearby neighbors in a single dimension. A training data set is utilized, and each point (indexed by i) within the n-dimensional set is assigned a probability estimate via RBFs for within class probability (P_{hit}) and between class probability (P_{miss}) (Eqs. 2, 3).

$$P_{i,\text{hit}} = \frac{\sum_{j=1}^{N_{\text{hit}}} e^{-(\epsilon \|x_i - x_j\|)^2}}{N_{\text{hit}}} \tag{2}$$

$$P_{i,\text{miss}} = \frac{\sum_{k=1}^{N_{\text{miss}}} e^{-(\epsilon \|x_i - x_k\|)^2}}{N_{\text{miss}}}. \tag{3}$$

The bandwidth variable ϵ is used to scale the kernel radius given a standard deviation. Given the class-specific probability estimates for each data point, we compute the relative separability of each data point between its hit class and miss class. This requires computing the Kullback–Leibler (KL) divergence of each point using both probability estimates (Eq. 4).

$$W_{i,\text{rbf}} = P_{i,\text{hit}} \cdot \log\left(\frac{P_{i,\text{hit}}}{P_{i,\text{miss}}} \right). \tag{4}$$

Table 2 Mean segment count ± standard deviation and [mean segment duration] by task and skill level

Skill level	Peg transfer	Pattern cutting	Suturing
'Obvious novice'	30.5 ± 4.6 [260 ms]	61.9 ± 18.1 [130 ms]	41.7 ± 18.3 [203 ms]
'Obvious expert'	24.8 ± 1.3 [105 ms]	27.1 ± 4.5 [68 ms]	12.4 ± 3.0 [107 ms]

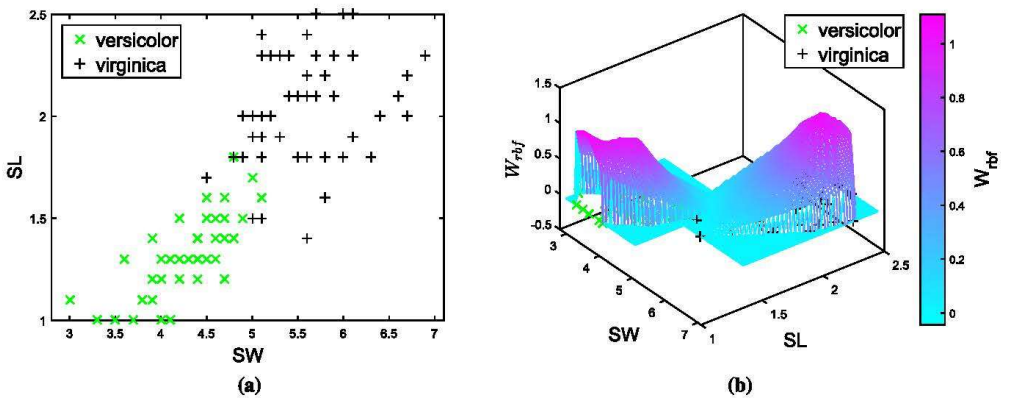


Fig. 1 RELIEF-RBF for sepal width (SW) and sepal length (SL) data from the Versicolor and Virginica classes. **a** 2D Fisher Iris data. **b** RELIEF-RBF weights

Each data point x_i in d-dimensional space ($d \leq n$) is assigned an estimate of separability $W_{i,\text{rbf}}$ (i.e., relevance in terms of classification use). The mean relevance weighting from all points in the training data set yields an aggregate estimate of the relevance weighting for that combination of features. This relevance weight is then compared with other combinations to improve feature selection for large, multidimensional, numerical data sets. A two-dimensional example of the relevance weights for two classes of the Fisher Iris data set [8] (Versicolor and Virginica) is given in Fig. 1. The RELIEF-RBF algorithm rewards only regions with high confidence of separability (high $W_{i,\text{rbf}}$), while penalizing *both* regions with a prevalence of all classes *and* regions that are data scarce (low $W_{i,\text{rbf}}$).

In both RELIEFF and RELIEF-RBF, all dimensions are mean–variance pre-scaled to account for data range effects. The weights for both methods are un-normalized and are used to compare the relative separability across dimensions.

Using both RELIEFF and the RELIEF-RBF, we investigated which states from the raw EDGE motion data had the highest separability. The states used in this study are given in Eq. (5) where $\dot{x}, \dot{y}, \dot{z}$ terms represent derivatives w.r.t. time of the Cartesian location of the surgical tool tip. χ_t is sample at each time step in the data set. The Cartesian position of the surgical tool $[x, y, z]$ was excluded because of its relationship to the present surgical gesture. All resulting feature combinations were investigated.

$$\chi_t = \left[\theta\dot{\theta}\ \dot{x}\ \dot{y}\ \dot{z}\ \ddot{x}\ \ddot{y}\ \ddot{z}\ \dddot{x}\ \dddot{y}\ \dddot{z}\ \|\dot{x},\ \dot{y},\ \dot{z}\|\|\ddot{x}, \ddot{y}, \ddot{z}\|\right]. \tag{5}$$

For comparison, we also applied RELIEF-RBF to the Fisher Iris data set, a well known, separable data set. Using the three surgical motion states with the highest RELIEF-RBF separability, we employed a random forest classification (100 trees) to examine the classification accuracy in a LOUO cross-validation scheme.

Intent vectors

We present a novel motion statistic for surgical skill classification. The 'intent vectors' statistic is based on the overall goal of a motion segment. Using the starting and ending location of a motion segment as endpoints, we compute a vector which represents the ultimate goal of that segment. We assume this intent vector is the ideal line of motion for a given segment; then we compute metrics which represent the amount of deviation from this optimal trajectory.

For a segment of Cartesian tool position data of length N, we have $\Psi = [D_1, D_2, \ldots, D_N]$ where $D_i = [x, y, z]$ represents the 3D location at time $t = i$. The intent vector is then computed in Eq. (6).

$$\vec{\text{IV}} = \frac{D_N - D_1}{\|D_N - D_1\|}. \tag{6}$$

From this intent vector, the progress of each point in Ψ along this line can contextualize other actions relative to the ultimate trajectory. The intent vector progress value (IVP) is computed according to Eq. (7) using a dot product operator and scaled by the magnitude of the intent vector (thus fixing the starting and ending points at 0 and 1). An illustrative example is given in Fig. 2a.

$$\text{IVP}_i = \frac{(D_i - D_1) \cdot \vec{\text{IV}}}{\|D_N - D_1\|}. \tag{7}$$

From the intent vector framework, we also compute the intent vector angle (IVA): the angle of motion relative to the

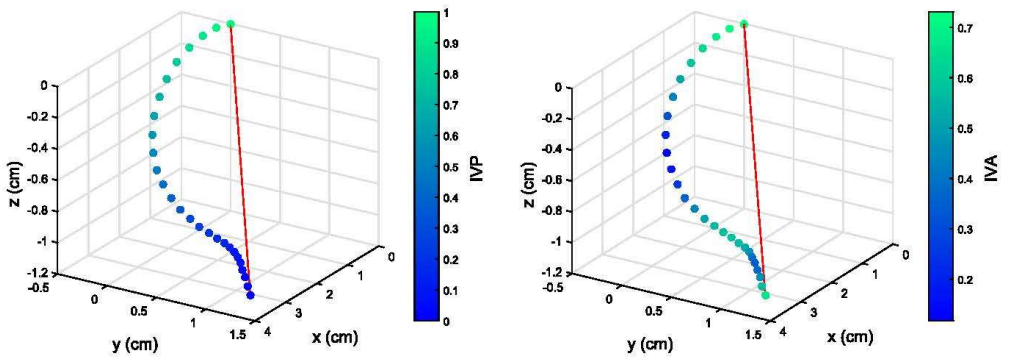


Fig. 2 Intent vector measures. **a** Intent vector progress in 3D. **b** Intent vector angle in 3D

overall angle of the intent vector. IVA is computed for each point in Ψ by taking the difference at a given point in time between the current tool location and the previous location $(D_i - D_{i-1})$ which is then normalized to give a unit vector in 3D space (S_i). Given this instantaneous unit vector, we compare with the overall intention, indicating the degree to which the tool is moving in the correct direction or doubling back (Eqs. 8, 9).

$$S_i = \frac{D_i - D_{i-1}}{\|D_i - D_{i-1}\|} \tag{8}$$

$$\text{IVA}_i = \cos^{-1}(S_i \cdot \vec{\text{IV}}). \tag{9}$$

The value of IVA is bounded between $0 < \text{IVA} < \pi$ since we are not concerned with the direction that the angle differs from the overall intent. An illustrative example is given in Fig. 2b. The intent vector framework was implemented for all motion segments within the EDGE data set. For each task, the IVA and IVP measures were compiled into a 2D feature vector with corresponding skill labels. A plot of IVA and IVP for the suturing task can be found in Fig. 4a.

Given the high-degree of similarity in the intent vector space, to use the intent vector data within a classification scheme we employed a classification approach which focuses on deviations from the region of high expert probability. We first identified the region in 2D IVA–IVP space with the highest density of expert surgical motion. We employed a modified version of the RELIEF-RBF algorithm and threshold the relevance weights for the expert class (Eq. 10).

$$W_{i,\text{exp}} = P_{i,\text{exp}} \cdot \log\left(\frac{P_{i,\text{exp}}}{P_{i,\text{nov}}}\right). \tag{10}$$

Here $W_{\text{exp}} = W_{\text{rbf}}$ from Eq. (4) where expert is the hit class. All training data are assigned a relevance weight relative to the expert data. A threshold on $W_{i,\text{exp}}$ is computed using an information gain maximization similar to the typical decision stump algorithm [10]. We identify a threshold (T_{w}) such that classification of the intent vector data follows Eq. (11) and maximizes the information gain $(\text{IG} = H(Y|X) - H(Y))$ for classification $(Y = \text{skilllevel})$ given $(X = [\text{IVA}, \text{IVP}])$.

$$Y = \begin{cases} \text{Novice}, & W_{\text{exp}}(X) < T_{\text{w}} \\ \text{Expert}, & W_{\text{exp}}(X) \geq T_{\text{w}}. \end{cases} \tag{11}$$

Using the relevance weight threshold, we retain all expert data in [IVA, IVP] space above T_{w} as 'true expert data' and train a Gaussian probability model for online classification $(P_{\text{exp}}(X|\mu, \sigma))$. A threshold value for this Gaussian model (T_{p}) is found by taking the $P_{\text{exp}}(X)$ at the minimum $W_{i,\text{exp}}(X) > T_{\text{w}}$ value.

The next step is to classify each individual time-indexed data point within a given segment for a specific surgeon. For surgeon (g) and segment (s), the time series data are given as $\Lambda_{g,s} = [\lambda_1, \lambda_2, \ldots, \lambda_N]$ where $\lambda_i = [\text{IVA}, \text{IVP}]$ at time $t = i$. Using $P_{\text{exp}}(X|\mu, \sigma)$, we classify each data point as 1 or 0 to signify novice or expert, respectively (Eq. 12). Values where $y_i = 1$ are considered a 'demerit' for behaving like a novice and are used in the overall evaluation of the motion.

$$y_i = \begin{cases} 1, & P_{\text{exp}}(\lambda_i) < T_{\text{p}} \\ 0, & P_{\text{exp}}(\lambda_i) \geq T_{\text{p}}. \end{cases} \tag{12}$$

Given a vector of time-indexed motion demerits $q_{g,s} = [y_1, y_2, \ldots, y_N]$, we compute a mean score for that particular segment $\text{SK}_{g,s} = \text{mean}(q_{g,s})$. Given the 1, 0 labels, this score has the effect of being very low for frequent expert motions and higher if motions fall outside the 'true expert'

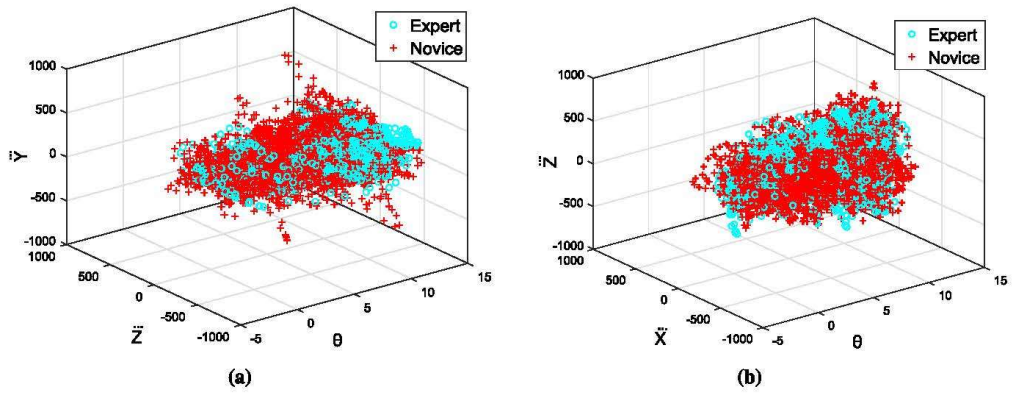


Fig. 3 Relevance weightings for raw motion states. **a** Top three RELIEFF states. **b** Top three RELIEF-RBF states

model (many novice demerits). We train a threshold based on the average SK scores (T_{sk}) for expert and novice surgeons using a decision stump approach. We employ a LOUO scheme per skill group (LOUOpG) (i.e., leave one obvious novice and one obvious expert out per training) and test each left-out surgeon based on all motion segments (Eq. 13).

$$C_g = \begin{cases} \text{Novice}, & \text{mean}(\text{SK}_{g,s}) > T_{sk} \\ \text{Expert}, & \text{mean}(\text{SK}_{g,s}) \leq T_{sk}. \end{cases} \tag{13}$$

For each LOUOpG iteration, we recompute all relevant measures and thresholds, i.e., W_{exp}, T_w, T_p, and T_{sk} based on the training data set alone, therefore limiting overfitting for the validation data.

In order to compare the accuracy of our classification approach, we utilized previously validated aggregate task metrics as highlighted in [5]. For this comparison, we used a feature vector comprised of tool path length, economy of motion (Eq. 14), motion smoothness, and motion curvature (Eq. 15, where $\dot{r} = \|\dot{x}, \dot{y}, \dot{z}\|$) ($\vec{\chi} = [\text{PL, EOM, MS, MC}]$). A linear discriminant analysis (LDA) classifier (class-based means and covariances, equal weighting) was trained on this feature vector to classify skill levels. We again employed a LOUOpG cross-validation with this classifier. We also examined classification using a combination of intent vectors and aggregate metrics with combined feature vector $\hat{\chi} = [\vec{\chi}, \text{mean}(\text{SK}_{g,s})]$. Again we utilized a standard LDA classifier in a LOUOpG cross-validation to classify a complete task.

$$\text{EOM} = \frac{\text{Path Length}}{\text{Task Time}} \tag{14}$$

$$\text{MC} = \frac{\dot{r} \times \ddot{r}}{|\dot{r}|^3}. \tag{15}$$

Results

Value of 'raw motion data' for classification

The relevance of the raw motion states was examined for all states in Eq. (5). The three motion states with the highest relevance weights according to RELIEFF were found to be $[\theta, \dddot{z}, \dddot{y}]$. The corresponding RELIEFF weights were $[2.3 \times 10^{-3}, 2.7 \times 10^{-3}, 3.0 \times 10^{-3}]$. A plot of these three states is given in Fig. 3a.

RELIEF-RBF gave slightly different states with high relevance. The motion states with the highest relevance weights according to RELIEF-RBF were found to be $[\theta, \dddot{x}, \dddot{z}]$. The corresponding RELIEF-RBF weight was 6.7×10^{-3} for this combination of states. A plot of these three states is shown in Fig. 3b. The additional relevance weights for the other motion states are not included for the sake of brevity but were all similarly low.

All states in the motion data had separability measures that were orders of magnitude lower than the separability of the Fisher Iris data set, which has a maximum relevance weight of 0.63 for sepal width and sepal length (RELIEF-RBF). Using a random forest classifier on the top RELIEF-RBF motion states gave a classification accuracy of 70.5% and an out-of-bag error of 0.28. Given the relatively low feature weights for the raw motion data, the resulting classification accuracy did not fulfill the MAC criterion, being well below 100%.

Intent vectors

A sample plot of the intent vectors space is given in Fig. 4a. These data indicate clear differences between novices and experts. Novices spend far more time outside the 0–1 range of the IVP, meaning they often backtrack and overshoot the starting and ending points. Additionally, experts spend a lot

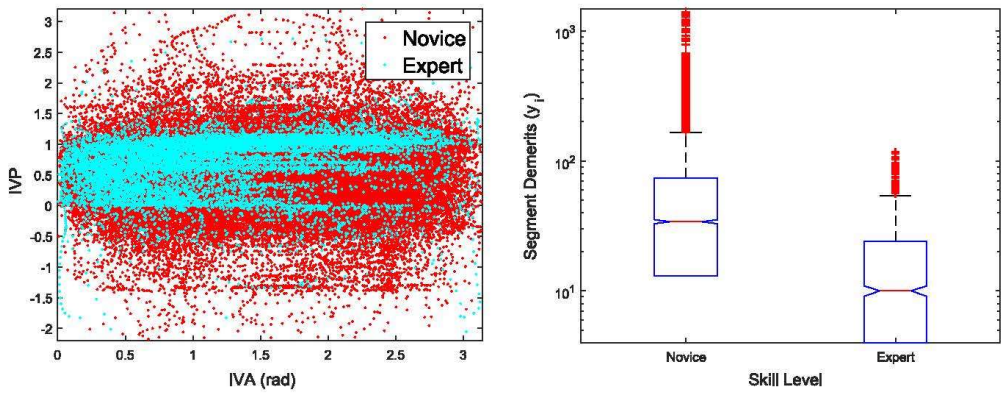


Fig. 4 Intent vector data (**a**) and demerit counts (**b**) (obvious novice and expert) for suturing task box-plot notch indicate range of 95% confidence for median separation. **a** IVA versus IVP with class labels. **b** Per-segment demerits (y_i)

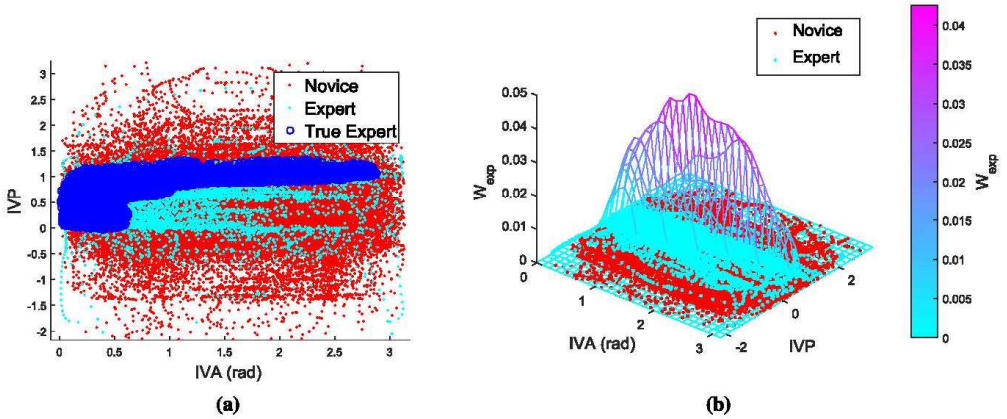


Fig. 5 Intent vector data with 'true expert' data and RELIEF-RBF weights (obvious novice and expert). **a** 'True expert' region. **b** RELIEF-RBF weights

of time with low IVA values meaning they generally head in the correct direction. However, experts also have varied IVA values around the endpoint of segments (IVP = 1), meaning that near the endpoint, experts make fine adjustments to their approach.

The intent vector classification yielded a large separation among segment demerit counts (y_i) between expert and novice surgeons. A plot of these values for each class is given in Fig. 4b. The mean segment demerit count was found to be 65.9 (std = 105.2) for novices and 22.6 (std = 27.7) for experts. The relevance weights (W_{exp}) and 'true expert' data in the intent vector space are shown in Fig. 5.

The intent vector framework yielded an average classification accuracy of 97% between novices and experts using a LOUOpG scheme for all tasks combined (Table 3). The intent vector approach fails to pass the MAC criterion for all tasks. However, it does achieve the MAC for the pattern cutting task.

An example plot of expert versus novice total segment demerits and the learned thresholds T_{sk} (Eq. 13) from all LOUOpG iterations is given in Fig. 6 for the intracorporeal suturing task. Results suggest the existence of an ideal threshold (obtainable using all available data) that provides clear separation between novice and expert data in the suturing task.

For comparison, the LDA classifier using the aggregate task metric features ($\bar{\chi}$) achieved the classification rates in square brackets in Table 3. These measures failed to achieve

Table 3 Intent vectors [aggregate metrics] {combined features} classification accuracy (%)

Skill level	Peg transfer	Pattern cutting	Intracorporeal suturing
Novice	96.5 [100[a]] {100[a]}	100[a] [96] {96}	100[a] [92.3] {92.3}
Expert	83.3 [83.3] {86.2}	100[a] [90] {100[a]}	92.3 [87.5] {100[a]}
Macro-accuracy	94.2 [97.1] {97.6}	100[a] [94] {97.2}	97.1 [90] {95.2}

[a] Achieves MAC criterion

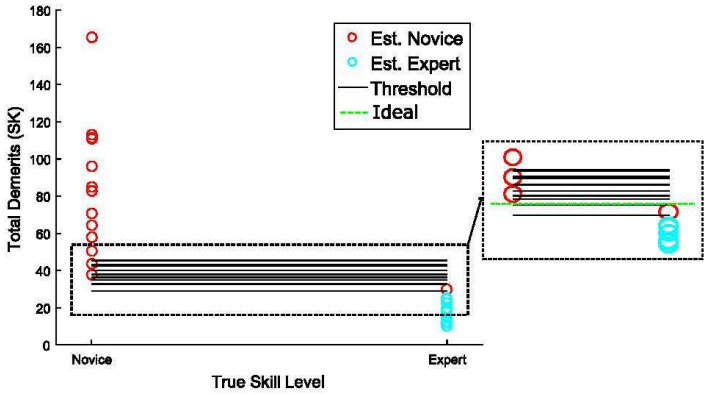


Fig. 6 LOUOpG classification using intent vectors with thresholds (T_{sk}) and ideal separable threshold

100% (macro-accuracy) classification for any of the tasks. The intent vector approach performed better than aggregate measures for both the suturing and cutting tasks, but worse in the peg transfer task. The combined feature vector $\hat{\chi}$ achieved equivalent or better macro-accuracy than the aggregate metrics alone for all tasks, indicating improved performance through the incorporation of intent vectors.

Conclusion

We presented the minimally acceptable classification (MAC) criterion for surgical skill classifiers. That is, given obvious expert and obvious novice data, a classification accuracy of 100% must be demonstrable as a minimal criteria for surgical skill classification. This requires stating both the classifier performance under LOUO-level cross-validation and enumerating its useful benefits over existing methods like summary metrics (e.g., task time).

We investigated the separability of raw tool motion data between obvious novices and experts with this MAC criteria in mind. As visible in Fig. 3, our results indicate extremely low separability—orders of magnitude lower than, say, the Fisher Iris dataset. This was true using both the RELIEFF and RELIEF-RBF feature selection algorithms. This suggests that motion data alone are statistically inseparable for classification given the MAC criterion. This is reiterated by the poor performance of the random forest classifier using raw tool motion alone. This motivates the inclusion of additional context (e.g., video data, tracking tissues, and tool–tissue interaction) to amplify the relevance of input data to the classification problem.

The intent vector feature and classifier performed surprisingly well given the observed low separability of the raw tool motion data. The overall classification rate of 97% rivals or surpasses prior the literature especially under LOUOpG cross-validation. We note that this approach fails to achieve the MAC criterion for all three FLS tasks. However, our intent vector classifier does partially succeed under the MAC criterion for two special cases: the cutting task and identifying obvious novices in the suturing task. Closer inspection in Fig. 6 reveals that the intent vector can fully separate the suturing task (and hence classify with 100% accuracy to achieve the MAC criterion) given an ideal threshold. This approach achieves equivalent or better results when compared with aggregate task metrics common in prior art. When used in the combined feature vector $\hat{\chi}$, we found that intent vectors improve classification accuracy when compared with the aggregate task metrics alone. Furthermore, for the cutting and suturing tasks, the intent vector provides additional value beyond summary metrics like task time. Notably, it returns classification results upon completion of each motion segment. This permits use cases such as (1) identifying only the worst portions of a surgical video for streamlined targeted review or (2) providing skill feedback in near real time at the completion of every motion. The segmentation approach used has the additional benefits of not requiring manual segmentation and being task agnostic.

We propose that the MAC criterion be adopted in surgical skill research as a minimal benchmark for a surgical skill classifier. Otherwise, the cost or complexity of sophisticated algorithms may not be justified. Using MAC also demands more carefully chosen ground truth skill categories to ensure accurate establishment of the ground truth, e.g., combining multiple criteria such as OSATS review, caseload, and procedural metrics. Failure to establish such a clean ground truth may hamper scientific progress in skill evaluation research.

This study has multiple limitations. This approach has only been applied to manual laparoscopic data on simulated tasks. Our conclusions may not hold for other contexts such as live surgery or robotic systems. The high selectivity of our 'obvious expert' inclusion criteria resulted in relatively small numbers of trials for cross-validation. Future work will include additional data collection to remedy this and applying the intent vector framework to ternary skill level classification. Additional analysis will investigate the concordance of intent vector metrics with FLS scores. We intend to compare our approach with the DCC and ribbon area measures [1,11]. This method has only been applied within our ballistic approach segmentation scheme; future work will investigate whether intent vectors can be applied to other actions such as needle passing. The current framework assumes the overall intent of each segment is correct and does not account for motion with incorrect intent. This segmentation scheme has the potential for false positives but is assumed to affect skill groups equally.

Acknowledgements R. Dockter was supported by the University of Minnesota Interdisciplinary Doctoral and Informatics Institute (UMII) MnDRIVE fellowships.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the EDGE study.

References

1. Ahmidi N, Gao Y, Béjar B, Vedula SS, Khudanpur S, Vidal R, Hager GD (2013) String motif-based description of tool motion for detecting skill and gestures in robotic surgery. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 26–33
2. Ahmidi N, Poddar P, Jones JD, Vedula SS, Ishii L, Hager GD, Ishii M (2015) Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. Int J Comput Assist Radiol Surg 10(6):981–991
3. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
4. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, Lendvay T (2014) Crowd sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 187(1):65–71
5. Chmarra MK, Klein S, de Winter JC, Jansen FW, Dankelman J (2010) Objective classification of residents based on their psychomotor laparoscopic skills. Surg Endosc 24(5):1031–1039
6. Datta V, Mackay S, Mandalia M, Darzi A (2001) The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. J Am Coll Surg 193(5):479–485
7. Faulkner H, Regehr G, Martin J, Reznick R (1996) Validation of an objective structured assessment of technical skill for surgical residents. Acad Med 71(12):1363–1365
8. Fisher RA (1936) The use of multiple measurements in taxonomic problems. Ann Eugen 7(2):179–188
9. Gomez ED, Aggarwal R, McMahan W, Bark K, Kuchenbecker KJ (2016) Objective assessment of robotic surgical skill using instrument contact vibrations. Surg Endosc 30(4):1419–1431
10. Iba W, Langley P (1992) Induction of one-level decision trees. In: Proceedings of the ninth international conference on machine learning, pp 233–240
11. Jog A, Itkowitz B, Liu M, DiMaio S, Hager G, Curet M, Kumar R (2011) Towards integrating task information in skills assessment for dexterous tasks in surgery and simulation. In: 2011 IEEE international conference on robotics and automation (ICRA). IEEE, pp 5273–5278
12. Kononenko I, Šimec E, Robnik-Šikonja M (1997) Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl Intell 7(1):39–55
13. Kowalewski TM, White LW, Lendvay TS, Jiang IS, Sweet R, Wright A, Hannaford B, Sinanan MN (2014) Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology. J Surg Res 192(2):329–338
14. Kowalewski TM, Sweet R, Lendvay TS, Menhadji A, Averch T, Box G, Brand T, Ferrandino M, Kaouk J, Knudsen B, Landman J, Leek B, Schwartz BF, McDougall E (2016) Validation of the AUA BLUS tasks. J Urol 195(4):998–1005
15. Lin HC, Shafran I, Yuh D, Hager GD (2006) Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. Comput Aided Surg 11(5):220–230
16. Malpani A, Vedula SS, Chen CCG, Hager GD (2014) Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In: International conference on information processing in computer-assisted interventions. Springer, pp 138–147
17. Reiley CE, Hager GD (2009) Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: Medical image computing and computer-assisted intervention—MICCAI 2009. Springer, pp 435–442
18. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM (2010) Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room a randomized controlled trial. Am J Surg 199(1):115–120
19. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R (2012) Sparse hidden markov models for surgical gesture classification and skill evaluation. In: International conference on information processing in computer-assisted interventions. Springer, pp 167–177

Laparoscopic Skill Classification using the Two-Third Power Law and the Isogony Principle

Anna French[1]
Timothy M. Kowalewski[1], **PhD**

[1]Department of Mechanical Engineering
University of Minnesota

1 Background

Surgical skill evaluation is a field that attempts to improve patient outcomes by accurately assessing surgeon proficiency. An important application of the information gathered from skill evaluation is providing feedback to the surgeon on their performance. The most commonly utilized methods for judging skill all depend on some type of human intervention. Expert panels are considered the gold standard for skill evaluation, but are cost prohibitive and often take weeks or months to deliver scores. The Fundamentals of Laparoscopic Surgery (FLS) is a widely adopted surgical training regime. Its scoring method is based on task time and number of task-specific errors, which currently requires a human proctor to calculate. This scoring method requires prior information on the distribution of scores among skill levels, which creates a problem any time a new training module or technique is introduced. These scores are not normally provided while training for the FLS skills test, and [1] has shown that FLS scoring does not lend any additional information over sorting skill levels based on task time. Crowd sourced methods such as those in [2] have also been used to provide feedback and have shown concordance with patient outcomes, however it still takes a few hours to generate scores after a training session.

It is desired to find an assessment method that can deliver a score immediately following a training module (or even in real time) and depends neither on human intervention nor on task-specific probability distributions. It is hypothesized that isogony-based surgical tool motion analysis discerns surgical skill level independent of task time.

2 Methods

2.1 Data Set

This study used tool motion data gathered from the EDGE (Electronic Data Generation and Evaluation) study [3]. This dataset contains 295 different samples of surgeons at varying skill levels interacting with a dry-lab surgical training environment performing 108 peg transfer (PegTx), 63 suturing and 124 circle cutting tasks.

Each sample is composed of a video recording of the training module (30 Hz), the cartesian space laparoscopic tool motion data corresponding to each video frame (30 Hz), features such as task name, and an FLS skill rating. The tool tip position and velocity measurements from the tool motion data were used to calculate our features of interest for evaluating skill. The FLS score was mapped in [1] to a ternary ranking of subjects as novice, intermediate or expert. The trials utilized for this experiment include 157 FLS novices, 71 FLS intermediates and 67 experts.

2.2 Analysis Methods and Algorithms

The time-agnostic velocity gain factor γ has shown promising results. In [5], this feature was used along with the two-third power law and the Isogony Principle to relate pencil tip velocity to the radius of curvature of 2D shapes sketched by a subject as

$$v(t) = \gamma(t)k(t)^{1/3}$$

where v is the velocity of the tool tip and k is the euclidean curvature (i.e. the instantaneous radius of curvature of the tool path.)

The mean and standard deviation of γ for the left and right hand over the course of the training run was taken to create the 4 features for each trial $\sigma(\gamma_L), \mu(\gamma_L), \sigma(\gamma_R)$ and $\mu(\gamma_R)$. These features were used to train a state vector machine (SVM) to predict the skill level of each trial.

The accuracy of the model was evaluated based on its agreement with the FLS classification for the trial, i.e. whether each model correctly classified the trial's FLS score grouping. Trials were grouped as either novice, intermediate, or expert. The FLS score groupings are used as the ground truth data.

To test the expected accuracy of an SVM which uses $\sigma(\gamma_L), \mu(\gamma_L), \sigma(\gamma_R)$ and/or $\mu(\gamma_R)$ as features, a 10-fold cross validation was performed. Cross validation helps ensure that the data used for testing was not included in the model training, and thus did not bias the model. The model's score grouping prediction was compared to the FLS score grouping from the EDGE data set. Each fold of the 10-fold cross validation samples trials evenly from the three skill levels. In addition, the 10-fold cross validation was performed 10 times (creating new 10-fold sets each time) in order to average out any fluctuations in accuracy due to the particular samples chosen for each 10-fold. This means, 100 different models comprised of 100 different partitions of the data were generated, and the statistics for the accuracies of each model are communicated in the box plot figures. In this study, we were only testing discrimination between novice and expert skill levels.

In each of the box plots, the 25th and 75th percentiles are displayed as the box boundaries while the median is the central line in each box. The whiskers mark the most extreme non-outlier points, and the + points are outliers.

3 Results

Task specific models were generated as a basis for comparison to other task-specific methods. In Figure 1, the classification

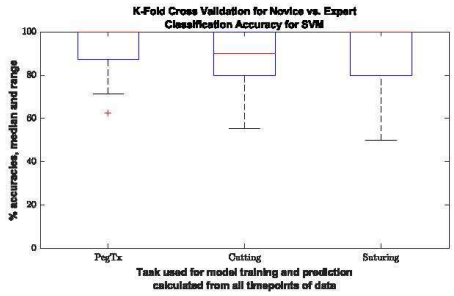


FIGURE 1: Prediction accuracy results of two-class classification, task specific trained model.

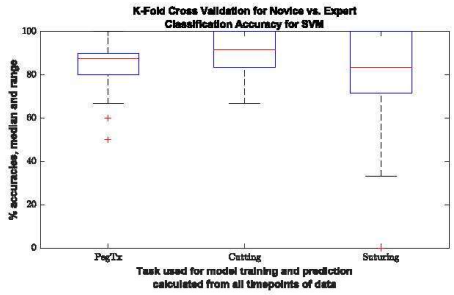


FIGURE 2: Prediction accuracy results of two-class classification, model trained in task-agnostic manner.

accuracies of task-specific models are displayed. To generate this figure, trials were separated based on which task was performed and a model to represent skill for each specific task was generated. A 10-fold cross-validation was done on each of these 3 partitions. This figure shows that for PegTx and Suturing models had a median classification accuracy of 100% per model.

In Figure 2, one model was trained using the full dataset in a task-agnostic manner (using trials across all tasks), and a similar cross validation technique to Figure 1 was performed. This figure shows good prediction accuracy for all three tasks despite the fact that the model was trained task-agnostic.

In Figure 3, five different models were trained using the full dataset in a task-agnostic manner, four of which were trained based on a single feature. The "all" column of this figure uses the same model from Figure 2. This shows the relative strength of each feature in prediction accuracy, where the accuracy of models trained using $\sigma(\gamma_R)$ or $\sigma(\gamma_L)$ alone are close in accuracy to the models trained using all four features.

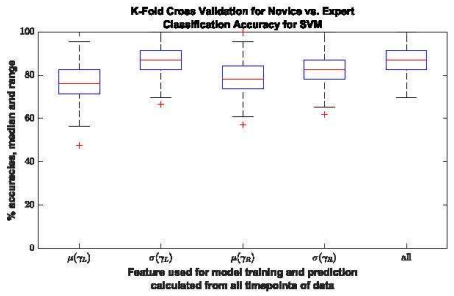


FIGURE 3: Prediction accuracy results of two-class classification, models trained using either single or all four features.

4 Interpretation

It was shown that task and time agnostic isogony-based features can be used to train automated skill evaluation models with good agreement to rough groupings of FLS scores. The ability to automate skill evaluation will allow surgeons in training to obtain feedback faster and more frequently. Further work will incorporate more trials to train more consistent models. These γ features will also be applied to other machine learning algorithms and used for 3-class classification to discern all FLS skill levels.

REFERENCES

[1] Kowalewski, T. M., 2012. "Real-time quantitative assessment of surgical skill". PhD thesis, University of Washington.

[2] Kowalewski, T. M., Comstock, B., Sweet, R., Schaffhausen, C., Menhadji, A., Averch, T., Box, G., Brand, T., Ferrandino, M., Kaouk, J., et al., 2016. "Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks". The Journal of urology, **195**(6), pp. 1859–1865.

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

[4] Rana, J., and Kowalewski, T., 2014. "Feasibility of a low-cost instrumented trocar for universal surgical procedure analyses". Journal of Medical Devices, **8**(3), p. 030936.

[5] Lacquaniti, F., Terzuolo, C., and Viviani, P., 1983. "The law relating the kinematic and figural aspects of drawing movements". Acta psychologica, **54**(1-3), pp. 115–130.