Student: Scott Pourshalchi

Mentors: Dr. Jeremy Brown, Dr. Anand Malpani, Dr. Gina Adrales

Course: 601.446 Computer Integrated Surgery II, Spring 2018

Seminar Paper Review

Citation:

J. D. Brown, C. E. O'Brien, S. C. Leung, K. R. Dumon, D. I. Lee and K. J. Kuchenbecker, "Using Contact Forces and Robot Arm Accelerations to Automatically Rate Surgeon Skill at Peg Transfer," in *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 9, pp. 2263-2275, Sept. 2017.

Project Recap:

The goal of this project is to develop an intelligent system that can objectively assess robotic surgical skill using performance data about how surgeons move their hands, connected instruments, and how the instruments interact with the surgical workspace. This will be accomplished by building a hardware and software platform that collects motion data from da Vinci and physical interaction data (forces on task board and accelerations of tool). This will combine two previously developed surgical skill assessment platforms. Our platform will then be used to collect pilot data from users of various robotic surgical skill levels. We will search for patterns in the data to prepare for its use in machine learning applications. Finally, we will write an IRB proposal to begin large scale data collection.

Paper Selection:

This paper was chosen to review as it was the first material our mentors suggested we read when beginning this project. The paper summarizes work completed by our mentor Dr. Jeremy Brown during his time at the University of Pennsylvania. It is a useful resource for our project as it describes hardware used in the data acquisition system we will be creating, it describes the data processing and features used in their machine learning techniques, and details a study similar to what we will create for our IRB proposal.

Problem Statement and Key Results:

The main problem described in this paper is the current method of skill assessment for robotic surgery. This method relies almost exclusively on structured human grading called global evaluative assessment of robotic skills (G.E.A.R.S.) which can be subjective, tedious, time consuming, and cost ineffective as raters are practicing physicians. However, according to the experiments described in the paper, a surgeon's skill at robotic peg transfer can be reliably

rated via regression using features gathered from force, acceleration, and time sensors external to the robot.

Significance:

These findings have significant impacts on the medical community. Firstly, implementation of automated skill assessment reduces the need for human raters to assess basic psychomotor skill development. This will save time, money, and may provide a more accurate assessment of skill. Additionally, this paper is one of the first published to demonstrate automatic skill assessment for robotic minimally invasive surgery via physical interaction information (external to robot). The use of physical interaction information gives this method advantages over methods that use robot motion. Finally, real-time feedback for a trainee learning robotic surgery through automatic skill ratings may allow trainees to learn faster.

Background:

Training with a clinical robot is the standard for training surgeons in robotic minimally invasive surgery. This method is preferred over virtual reality training because the process is closer to actual surgery than what is currently possible to simulate through VR. However, as mentioned skill evaluation is often subjective, tedious, time consuming, and cost ineffective. Previous work was published that demonstrated the use of robot kinematics to assess skill during training and actual surgical procedures. However, kinematics-based methods cannot account for potential master-slave misalignments due to sensor error or for unmeasured quantities such as compliance and mechanical wear. Thus, the use of physical interaction information was considered by the authors.

Few papers have measured the physical interaction between the robot and the environment when analyzing trainee skill development. Previous work by these authors showed that the root mean square of high frequency vibrations of both the robotic tools and forces exerted on the task materials are greater for novices than experts.[1] After obtaining these findings, the experiment described in this paper was conducted.

Data Collection:

The hardware used in this project includes 2 high bandwidth 3 axis accelerometer clips for the two primary robotic arms, 1 high bandwidth 3 axis accelerometer clip for the endoscope, and 1 “Smart Task Board” seen in figure 1. The accelerometers can be attached to the robot via 3D printed clips seen in figure 2. The “Smart Task Board” consists of a peg transfer task mounted on a platform with a three-axis force sensor. Data collection was coordinated with a Python script at 3 kHz. Video was recorded through an s-video connection.

[1] K. Bark et al., "Surgical instrument vibrations are a construct-valid measure of technical skill in robotic peg transfer and suturing tasks", *Proc. Hamlyn Symp. Med. Robot.*, pp. 50-51, 2012.

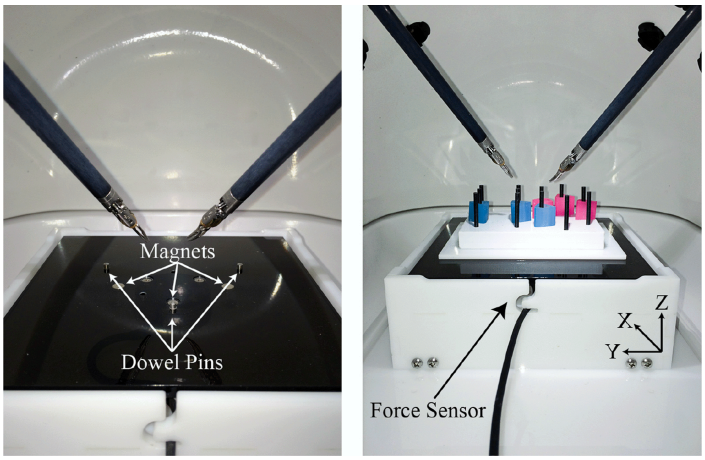


Figure 1: Peg Transfer Task

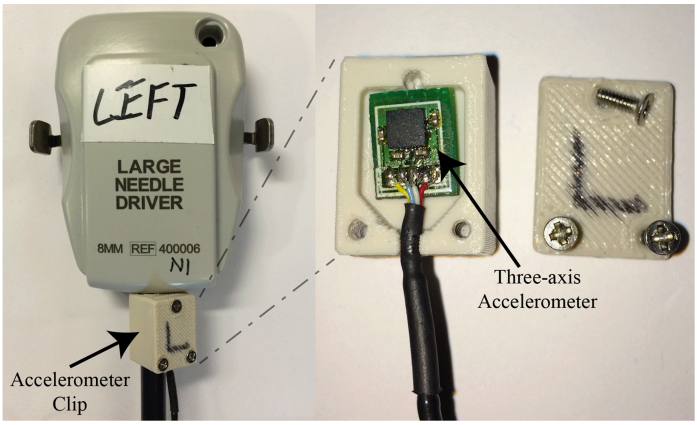


Figure 2: Accelerometer Clip

Participants:

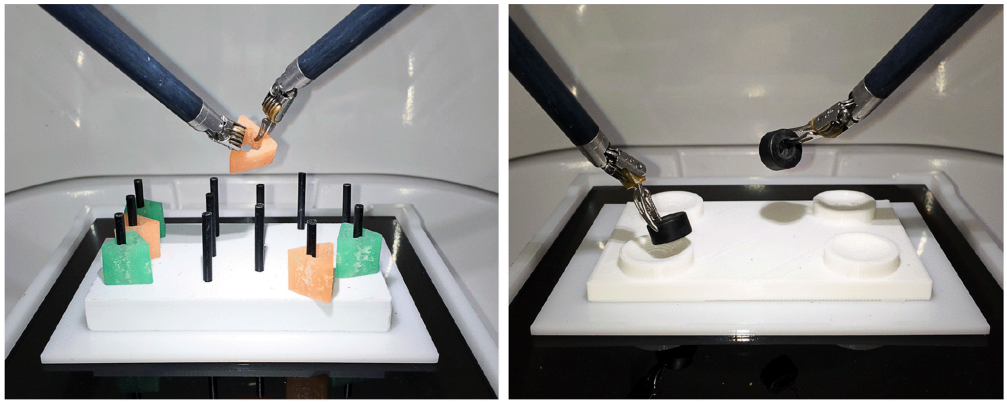
38 clinicians participated in this study. Participants were obtained from various skill levels in robotic surgery as shown by Table 1. Skill level labels were obtained from self-reported familiarity & number of robotic cases completed.

TABLE I
PARTICIPANT DEMOGRAPHICS

Handedness	Left	Right	Ambidextrous	
	3	32	3	
Familiaritywith Robot	None	Limited	Moderate	Extensive
	13	10	6	9
# Robotic Cases	None	1–100	101–500	501 +
	22	6	7	3

Each participant was allowed to warm up with a practice task (Figure 3, image b), then completed 3 trials of peg transfer (Figure 3, image a), and finally completed a demographic questionnaire. Trials were rated for skill by 2 surgeons with previous experience rating on the G.E.A.R.S. scale. The G.E.A.R.S. ratings contain 5 domains: Depth Perception, Bimanual Dexterity, Efficiency, Force Sensitivity, and Robotic Control. To ensure interrater reliability,

raters were given time to “calibrate”. This consisted of giving each rater a set of 10 diverse videos and time to discuss the ratings. The interrater reliability of ratings was assessed using intra-class correlation coefficient (ICC). 0.6 was chosen as the minimum acceptable ICC for “good” reliability. As shown in table 2, the total ICC for each domain was sufficient.



(a) (b)

Figure 3. a) Peg transfer task b) warm up task

TABLE II
FREQUENCIES OF GEARS RATINGS AVERAGED ACROSS RATERS AND ROUNDED, AND FINAL ICC FOR RATED TRIALS

GEARS Domain	Ratings					ICC
	1	2	3	4	5	
Depth Perception	0	14	45	41	10	0.76
Bimanual Dexterity	0	9	41	44	16	0.80
Efficiency	3	14	44	30	19	0.89
Force Sensitivity	0	15	42	43	10	0.74
Robotic Control	2	10	48	42	8	0.80
Overall						0.88

Machine Learning:

To prepare for machine learning method the time series data was broke down into a set of discrete features. Firstly, acceleration data used to calculate roll (rotation around the shaft) and pitch (shaft angle relative to the horizontal) given by:

$$Roll\ \phi = \tan^{-1}\frac{a_{fy}}{a_{fz}}$$

$$Pitch\ \theta = \tan^{-1}\frac{-a_{fx}}{\sqrt{a_{fy}^2 + a_{fz}^2}}$$

Time features included total elapsed time, total active time. This included the square root and log of these values as skill may be non-linear with time. Descriptive features included mean, standard deviation, minimum, maximum, range, Root Mean Square (RMS), Total Sum of Square (TSS), time integral of force directions and magnitude; tool and camera roll and pitch angles, angular velocity, accelerations; product of right and left tool acceleration in each frequency band; product of force magnitude and right/left tool acceleration in each frequency band

10 learners were created using both regression and classification for each G.E.A.R.S. domain. 33 participants were used for a training set and 4 participants were reserved for testing. This approximately matches the 90%|10% standard of machine learning literature. The regression learners were computed in MATLAB using LIBSVM library, Glmnet library, and the Statistics and Machine Learning Toolbox. Random forest classification learners were implemented TreeBagger function in MATLAB's statistics and machine learning toolbox. Training took approximately 30 min for all 5 domains and 30 sec for rating calculation/classification.

Results:

The most important features for each G.E.A.R.S. domain are identified in figure 4. It was determined that both regression and classification leaners could be used to assess surgical skill accurately but regression learners showed slight advantages. Table 3 highlights the exact accuracy of skill assessments. Additionally, precisions obtained were greater than 0.2 suggesting that performance was better than random chance. Finally, ICC was calculated to assess how well the learners served as raters of assessments objectively.

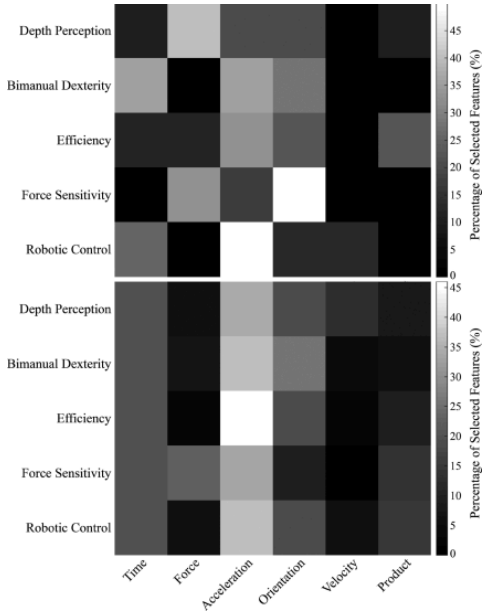


Figure 4: Relevant features for each G.E.A.R.S. domain

TABLE III
EXACT ACCURACY ACROSS TESTING SETS

GEARS Domain	Regression Learner	Classification Learner
Depth Perception	63.3 ± 9.5%	71.7 ± 9.5%
Bimanual Dexterity	66.7 ± 11.8%	53.3 ± 16.2%
Efficiency	73.3 ± 16.0%	58.3 ± 8.3%
Force Sensitivity	63.3 ± 9.5%	51.7 ± 10.9%
Robotic Control	71.7 ± 12.6%	75.0 ± 15.6%

Values shown are mean ± standard deviation across the five testing sets.

TABLE V
RANGE (MEDIAN) OF ICC(2,4) BETWEEN THE THREE RATERS AND EACH LEARNER (REGRESSION AND CLASSIFICATION) FOR THE FIVE RESERVED TESTING SETS

GEARS Domain	Regression Learner	Classification Learner
Depth Perception	0.80–0.88 (0.81)	0.76–0.84 (0.83)
Bimanual Dexterity	0.71–0.91 (0.86)	0.71–0.86 (0.84)
Efficiency	0.84–0.93 (0.88)	0.83–0.91 (0.88)
Force Sensitivity	0.70–0.90 (0.79)	0.66–0.84 (0.76)
Robotic Control	0.66–0.87 (0.79)	0.69-0.86 (0.81)
Overall	0.88–0.93 (0.89)	0.87–0.89 (0.89)

Relevance to Project:

This paper was extremely relevant to our project. It describes hardware information of the data acquisition system we will use (relates to minimum deliverables). It describes data preprocessing and important features used in machine learning techniques (relates to expected deliverables). Finally, it describes user study similar to what we will create for IRB proposal (relates to maximum deliverables). Additionally, the discussion suggests next steps for project is combination of physical and motion interaction data – relates directly to our project.

Pros and Cons:

There were several pros of this paper. Online supplements contained machine learning performance analysis without force data which can't be accessed in vivo. This supports the methods viability. Additionally, there is a lengthy discussion section which evaluates validity of results, impact of results, etc. Direct comparison of time saved through this approach (110 trials rated in 6 months by human grading, 20 min by regression) support the use of this approach. Finally, this skill assessment is much more flexible than those based on kinematics as it doesn't interfere with robot control or operation and accounts for master-slave misalignment and compliance.

However, some cons of the paper exist as well. More descriptions of why features were chosen and why the machine learning methods were chosen would have been helpful. Additionally, the lowest accuracy G.E.A.R.S. domain was force sensitivity which something explicitly measured. This suggests that they were possibly not examining the right features. The paper demonstrates results for peg transfer – actual surgery is much more complex and the skills are not necessarily equivalent. Finally, there was unequal representation among skill levels and low number of participants

Conclusion:

This paper is a great resource for this project. It explains the motivation, hardware, data processing, and software related to the project. We will continue to refer to this paper in the future.