Using Contact Forces and Robot Arm Accelerations to Automatically Rate Surgeon Skill at Peg Transfer

Jeremy D. Brown, *Member, IEEE*, Conor E. O'Brien, Sarah C. Leung, Kristoffel R. Dumon, David I. Lee, and Katherine J. Kuchenbecker, *Member, IEEE*

Abstract*—Objective: Most trainees begin learning robotic minimally invasive surgery by performing inanimate practice tasks with clinical robots such as the Intuitive Surgical da Vinci. Expert surgeons are commonly asked to evaluate these performances using standardized five-point rating scales, but doing such ratings is time consuming, tedious, and somewhat subjective. This paper presents an automatic skill evaluation system that analyzes only the contact force with the task materials, the broad-bandwidth accelerations of the robotic instruments and camera, and the task completion time. Methods: We recruited $N = 38$ participants of varying skill in robotic surgery to perform three trials of peg transfer with a da Vinci Standard robot instrumented with our Smart Task Board. After calibration, three individuals rated these trials on five domains of the Global Evaluative Assessment of Robotic Skill (GEARS) structured assessment tool, providing ground-truth labels for regression and classification machine learning algorithms that predict GEARS scores based on the recorded force, acceleration, and time signals. Results: Both machine learning approaches produced scores on the reserved testing sets that were in good to excellent agreement with the human raters, even when the force information was not considered. Furthermore, regression predicted GEARS scores more accurately and efficiently than classification. Conclusion: 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: We expect improved trainee learning as a result of providing these automatic skill ratings during inanimate task practice on a surgical robot.

Manuscript received August 10, 2016; revised November 4, 2016; accepted November 12, 2016. Date of publication December 7, 2016; date of current version August 18, 2017. This work was supported in part by an Intuitive Surgical Technology Research Grant and by a University of Pennsylvania Academic Postdoctoral Fellowship held by J. D. Brown. *Asterisk indicates corresponding author.*

*J. D. Brown is with the Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: brownjer@seas.upenn.edu).

C. E. O'Brien was with the Department of Mechanical Engineering and Applied Mechanics and the Department of Computer and Information Science, University of Pennsylvania. He is now with Otherlab.

S. C. Leung was with the Department of Mechanical Engineering and Applied Mechanics and the Department of Computer and Information Science, University of Pennsylvania. She is now with Boeing.

K. R. Dumon and D. I. Lee are with the University of Pennsylvania Hospital System.

K. J. Kuchenbecker is with the Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org(File Size 388 KB).

Digital Object Identifier 10.1109/TBME.2016.2634861

***Index Terms*—Machine learning, physical interaction, robotic surgery, skill evaluation.**

I. Introduction

BEGINNING with Halsted's use of the "See One, Do One, Teach One" philosophy in the first medical residency program [1], hands-on training has become an important aspect of all surgical curricula. While Halsted's apprenticeship model flourished for nearly a century, implementation of an 80-hour resident work week [2] and increased emphasis on patient safety [3] have forced a large portion of psychomotor skill development to now take place through simulation-based training outside the operating theater.

For minimally invasive surgery, simulation-based training must adapt the psychomotor skills a trainee has developed for traditional open surgery to a new surgical landscape in which long thin surgical instruments are manipulated through small incisions under the guidance of a lighted scope. The Fundamentals of Laparoscopic Surgery (FLS) curriculum helps train surgeons for laparoscopic procedures. However, no standardized training curriculum exists yet for robotic minimally invasive surgery (RMIS), which is commonly used in urologic surgery [4], gynecologic surgery [5], and general surgery [6]. Despite enhanced visualization and increased dexterity [7], [8], the lack of haptic feedback in RMIS requires trainees to learn to rely more heavily on vision [9].

Recent efforts such as the Fundamentals of Robotic Surgery (FRS) [10] are developing standardized training protocols for the Intuitive Surgical da Vinci RMIS system; the FRS employs simulation-based training in the form of virtual reality (VR) training, structured inanimate task training with the clinical robot, and *ex vivo* animal model training with the clinical robot. VR training has been shown to correlate well with inanimate and *in vivo* training [11], and it has some advantages over the other two methods: VR trainers are less expensive than the clinical robot, require no disposable resources, have no clinical scheduling constraints, and allow for structured task practice in a low-stakes controlled environment [3], [12]. Still, VR trainers are capable of delivering only a virtual approximation of the clinical robot and the task materials, which may not exactly match how the real system behaves. For example, robotic surgeons are known to compensate for the dynamics of

the robot hand controllers [13], [14]; desktop systems such as the dV-Trainer (Mimic Technologies Inc., Seattle, WA, USA) have control interface dynamics that are different from those of the clinical robot. In addition, while these trainers have face, content, construct, and concurrent validity [3], [12], predictive validity has been shown only for specific simulators with respect to completion time for *in vivo* animal model training [15], completion time and number of instrument movements for inanimate task training [16], and bimanual dexterity, efficiency, and instrument plus camera awareness for extremely novice trainees with *ex vivo* tissue [17].

Although it may be difficult at hospitals where all robots see high clinical use, training with a clinical robot is still considered the gold standard [11], [12]. Structured inanimate tasks are well suited for teaching many of the fundamental psychomotor skills such as camera movement, tool manipulation, and needle handling, and they have been shown to correlate more highly with *in vivo* training than VR training [11]. *In vivo* training, on the other hand, is more suited for practicing surgical procedural technique [11]. In both cases, the current method for skill evaluation is structured human grading, in which an expert observer watches the live or recorded training and provides oral or written feedback to the trainee. Although expert evaluation is key to the training process, it can be subjective, tedious, time consuming, and cost ineffective (as most expert raters are practicing physicians). While crowdsourcing the rating process was recently shown to be a viable alternative to expert ratings [18], its adoption in clinical settings has been limited.

Considerable research has analyzed surgical instrument motions to provide automatic objective feedback to a trainee practicing on the clinical robot. In particular, researchers have worked to decode the motion language of surgery at both the task level and the motion segment level [19]. These approaches originated with open and laparoscopic surgery [20]–[23] but have found their way into robotic surgery due to the copious data available from the robot. For example, Lin *et al.* developed a technique for parsing raw motion data from a four-throw suturing task performed on a da Vinci surgical robot into a labeled sequence of surgical gestures for skill evaluation [24]. Similarly, Ahmidi *et al.* worked to detect surgical gestures and skill based on descriptive curve coding and a common string model classification approach [25].

One benefit of these motion-based approaches is that they require access only to the robot kinematics, so they can be used to assess skill during both training and actual surgical procedures. At the same time, their basis on the robot's motions (measured by internal joint position sensors) means they cannot account for potential master–slave misalignments due to sensor error [26] or for unmeasured quantities such as compliance and mechanical wear in the surgical tools. They are also blind to the exertion of large forces or the occurrence of rough contacts with the task materials. Therefore, motion analysis cannot completely describe the robot's physical interactions. In an effort to overcome this limitation, researchers have begun performing motion analysis from the video stream from the robotic camera [27], [28]. These approaches, however, lag significantly behind traditional motion-based assessment due to the complexity of analyzing 6-D motion from a 2-D video.

As a complementary approach to motion-based skill assessment, we propose to use the physical interactions between the robotic tools and the training environment as a basis for skill evaluation, irrespective of the robot's motions. In support of this strategy, we previously showed that the root mean square (RMS) and/or total sum of squares (TSS) of both the high-frequency vibrations of the robotic surgical instruments and the forces exerted on the task materials are significantly greater for novices than for experts during peg transfer, needle passing, and suturing tasks [29], [30]. Tool vibrations and contact forces are thus construct-valid measures of RMIS skill. This result is consistent with a broad set of findings that physical interaction signals are important for assessing skill in endovascular catheterization [31], endoscopic sinus surgery [32], natural orifice translumenal endoscopic surgery [33], and laparoscopic surgery [34]–[36].

Despite this wealth of published evidence in related areas, few researchers have measured the physical interactions between the robot and the surgical environment when analyzing trainee skill development. The lack of haptic feedback in robotic surgery makes such investigations especially interesting and important. In particular, we believe the rich physical interactions that occur when a trainee brings the robot's tools into contact with the surgical environment can potentially indicate his or her skill across a range of domains. It may even be possible to predict skill metrics defined by a validated assessment tool like the Global Evaluative Assessment of Robotic Skill (GEARS) [37].

This paper reports a new technique that uses external sensors to measure the robot's physical interactions and automatically evaluate a trainee's technical skill at an inanimate task. We employ a supervised machine learning approach that predicts GEARS scores from the forces applied to the task materials, the broad-bandwidth accelerations of the surgical instruments and the camera, and completion time. We demonstrate our technique through application to robotic peg transfer. In what follows, we describe the details of the data collection apparatus, the human subject data collection procedures, and the development and evaluation of machine learning algorithms that automatically predict skill. We end by discussing the implications of this new approach and areas of future research. Interested readers can find more information about our system design, task selection, feature calculation, and data analysis in the supplemental document associated with this paper.

II. Data Collection

The first step in creating an automatic skill assessment tool for robotic surgery training is to record a dataset of tasks performed by surgeons at a wide range of skill levels. Specifically, we sought to capture the physical interactions between the surgical robot and the training task. This section details the custom hardware and the calculations needed to record and process the surgical task performance data. We then discuss the study that was conducted to obtain the data.

A. Hardware

We have developed a Smart Task Board (STB) to record the physical interaction data from the patient-side manipulators

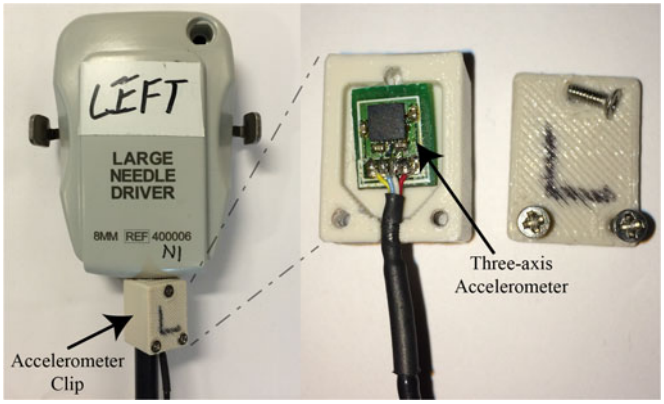

Fig. 1. Accelerometer clips consist of a custom 3-D-printed bracket that snaps onto the instrument or camera shaft, plus a three-axis accelerometer mounted to a custom circuit board.

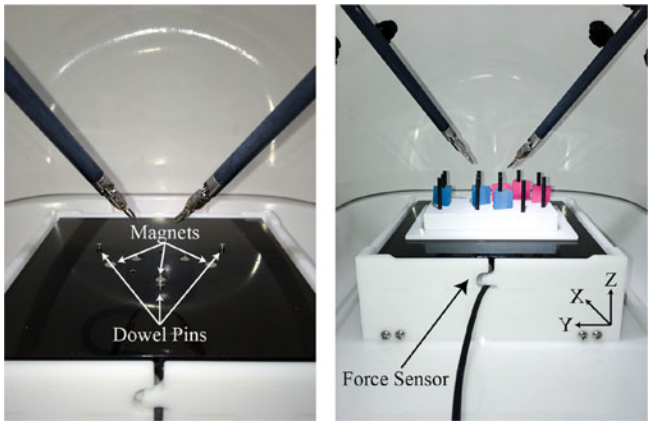

Fig. 2. Task platform with dowel pins and magnets for task board mounting, as well as a force sensor for measuring the forces that the surgeon applies to the task materials through the patient-side manipulators. The coordinate frame shows the orientation of the force sensor's axes.

of an Intuitive da Vinci surgical system. The STB consists of accelerometer clips for the two primary robotic arms, an accelerometer clip for the robotic camera arm, a task platform containing a three-axis force sensor, a custom signal conditioning box, and an Intel NUC computer for data acquisition.

The accelerometer clips are custom 3-D-printed brackets that snap onto the 8-mm instrument shafts and the 12-mm camera shaft. Each bracket contains a broad-bandwidth three-axis microelectromechanical-system-based accelerometer (LIS344ALH) mounted to a custom accelerometer circuit board, as shown in Fig. 1.

The task platform is a custom acrylic base that fits in the bottom of the white da Vinci skills dome. A raised platform is mounted on top of a three-axis force sensor (ATI Mini40 SI-40-2) at the center of the acrylic base. The top plate of the platform features dowel pins and magnets to ensure the task materials are mounted to the plate in a repeatable manner (see Fig. 2). The Mini40 signal conditioning box is inside an enclosure with a custom data acquisition board that features a Teensy 3.1 microcontroller with a 32-bit ARM Cortex microprocessor, as well as other chipsets for filtering, buffering, and analog-to-digital conversion.

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

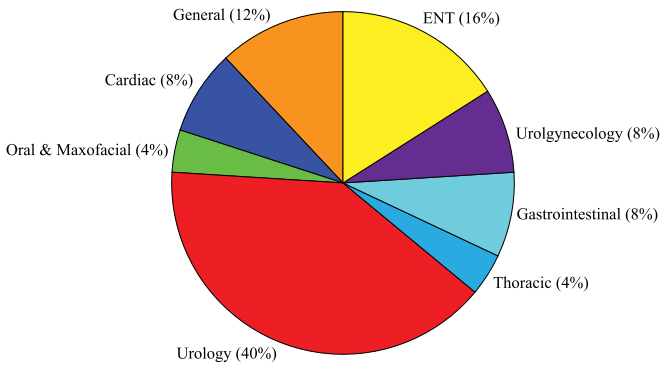

Fig. 3. Surgical specialty of resident, fellow, and attending participants

The nine single-ended analog inputs for the three three-axis accelerometers and the six differential analog inputs for the force sensor are sampled at 3 kHz and written to disk on the Intel NUC computer. In addition to recording the sensor data, the computer also records the live video feed from one of the robot cameras via an s-video connection. The entire data recording process is controlled through a Python script that zeros the force sensor before recording data. The user starts and stops the recording process using a foot pedal, so that he or she can already be holding the da Vinci masters when the trial starts. A strip of five lights mounted above the robot view port indicates the current recording status to the user.

B. Human Subject Study

We conducted a study to collect a large corpus of surgical skill performance data. The study was designed to capture performance metrics in five of the six domains measured by the GEARS evaluation tool [37]: Depth Perception, Bimanual Dexterity, Efficiency, Force Sensitivity, and Robotic Control. We omitted the GEARS domain on Autonomy because all participants were able to complete the task without verbal prompting. Participants were compensated for their participation with a $25 gift card. All study procedures were approved by the University of Pennsylvania Institutional Review Board under protocol #820759.

1) Participants: We tested $N = 38$ participants (22 males, 16 females, mean age of 31.5 $\pm$ 7.2 years) from a broad range of training levels, including 13 fourth-year medical students in the Agnew Surgical Society at the University of Pennsylvania Perelman Medical School, 14 surgical residents, four surgical fellows, and seven attending surgeons. Table I lists details of our participant population, and Fig. 3 shows the

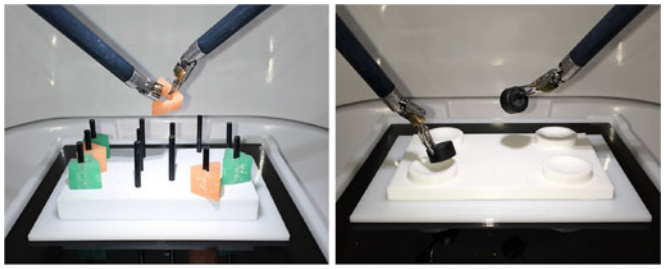

Fig. 4. (a) Peg transfer task and (b) practice task used to help familiarize participants with the da Vinci Standard robot platform.

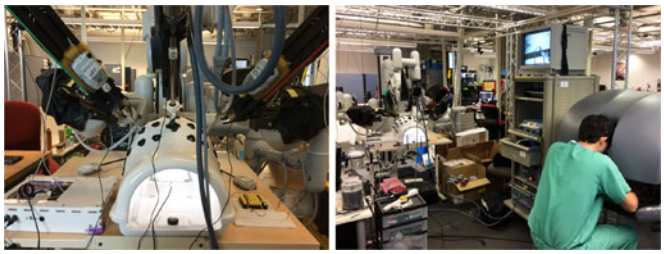

Fig. 5. Study setup. da Vinci patient-side cart with arms docked in the skills model shell according to a standard pelvic anatomy. The participant sits at the da Vinci master console to perform the task.

surgical specialties across our resident, fellow, and attending participants.

2) Experimental Setup: The study took place in the University of Pennsylvania's General Robotics, Automation, Sensing, and Perception Laboratory on a da Vinci Standard surgical system augmented with our STB data collection system. Participants used the da Vinci to perform the peg transfer task described below.

Peg transfer: Six triangular objects are placed on the left side of a pegboard. The participant picks up each object with their left tool, transfers the object midair to their right tool, and places the object on a peg on the right side of the board, as shown in Fig. 4(a). After transferring all six objects, the participant returns the objects to the pegs on the left side of the board by reversing the process. Participants were instructed to retrieve objects that fell on the taskboard with the tool from which it fell. If an object fell off the taskboard, participants were instructed not to try to retrieve it. Peg transfer was completed with two 8-mm-diameter EndoWrist Maryland Bipolar Forceps tools. This relatively simple task was adopted from the FLS manual skills test [38] and conducted in the da Vinci skills model shell (see Fig. 5).

3) Experimental Procedure: After giving informed consent, each participant sat at the da Vinci master console as shown in Fig. 5. The experimenter explained the da Vinci Standard system, including adjusting the ergonomics, focusing the camera, and clutching the tools and camera. Next, participants spent at least three minutes doing a warm-up task that featured four elevated podiums and two rubber objects that could be moved from one podium to another, as shown in Fig. 4(b). This practice time was required for all participants including experienced robotic surgeons because clutching is controlled somewhat differently between different da Vinci models.

After completing the practice session, the participant was shown how to operate the data recording system using the foot pedal and the light strip. Participants then viewed static images depicting the peg transfer task procedure and were instructed to complete it as well as possible. Participants were informed that the task was timed. Before the participant began the task, the tools were reloaded to reset their configuration, and the camera was adjusted to give a global view of the task board and the tool tips. The participant completed the peg transfer task three times, with the tools and camera reset after each repetition, and then completed two additional tasks that we will analyze in future work, followed by a demographic questionnaire. The entire testing session lasted about 60 min.

III. Data Preparation

We obtained accurate skill-level ratings (labels) for the recorded peg transfer trials, and we extracted a collection of discrete features from the time-series data. The steps described below use data from only 37 participants because all three trials from one participant (a fourth-year medical student) were corrupted while being saved. In addition, one trial from a separate participant (an experienced robotic surgeon) was not included because the robot encountered an error that forced a restart, and the participant was not able to repeat the trial due to time constraints. Thus, instead of 114 trials (three trials for each of the 38 participants), our data preparation was performed on the 110 trials that remained.

A. Surgical Skill Rating

We used the GEARS assessment tool [37] to obtain skill ratings for each recorded trial. Two expert robotic surgeons (>300 cases) with prior experience as GEARS raters were recruited to serve as raters: One is a urologic surgeon, and the other is a bariatric surgeon, both at the University of Pennsylvania Health System. One of the experimenters also served as a nonexpert rater.

Video rating took place using secure web-based surveys administered through Qualtrics. Each survey was eight to ten pages in length, with each page containing the embedded video from a unique de-identified trial along with the associated GEARS questions, which are rated on a five-point scale from 1 (lowest) to 5 (highest). Each survey contained a random selection of trials presented in random order. The raters never received any information about the subjects when performing the ratings.

To ensure good interrater reliability, the expert surgeon raters went through a calibration procedure. Each rater was given the same set of ten diverse videos to rate. Afterward, the raters met to discuss their ratings, giving special attention to the questions where their ratings differed by more than one point. In this manner, the raters were able to establish what level of observed performance corresponded to each rating. After this calibration, the expert surgeon raters rated a new set of ten diverse videos. The interrater reliability of these ratings was assessed using the intra-class correlation coefficient (ICC) [39].

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

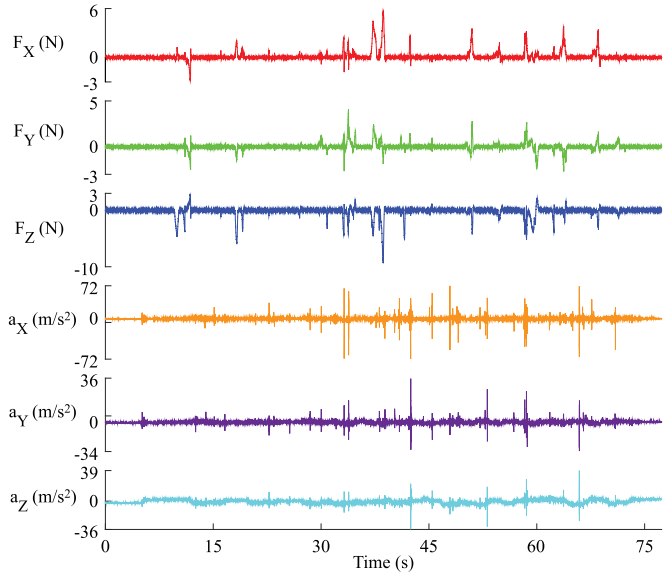

Fig. 6. These raw haptic signals were recorded by our system while one participant performed one trial of the peg transfer task. The top three traces are the X, Y, and Z components of the force signal. The bottom three traces are the X, Y, and Z components of the acceleration of the right tool. All three axes of the force signal show when the participant exerted force on the pegboard. Likewise, all three axes of the acceleration signal show when the right tool makes a hard contact with the other tool or the pegs on the pegboard; the DC acceleration values also show the tool's orientation relative to gravity.

ICC was calculated for each of the five GEARS questions, as well as overall. A value of 0.6 was chosen as the minimum acceptable ICC for "good" reliability. For any question with an ICC below 0.6, the raters reconvened to discuss possible discrepancies and update their ratings as appropriate. Once the ICC was above 0.6, the calibration was complete. The nonexpert rater conducted a similar calibration by comparing his or her ratings against the calibrated ratings by the expert surgeon raters. Table II shows the frequencies of the rounded average ratings produced by our three raters, as well as the ICC for all ratings for each domain and the overall summed GEARS score.

B. Feature Extraction

Our time-series data consisted of one three-axis force signal and three three-axis accelerometer signals. Fig. 6 shows the raw force signals and the accelerations of the right tool from one representative peg transfer trial. To develop a machine learning algorithm based on our surgical skill performance data, we broke this time-series data into a set of discrete features that describe the data. The readings from each of these accelerometers respond to two types of stimuli: the shaft's high-frequency vibrations, which mainly stem from contact with stiff objects, and the shaft's orientation relative to gravity [40]. We considered these sources to be potentially useful signals given results from our own vibration-based research [29], [30] and other motion-based assessments [24], [25].

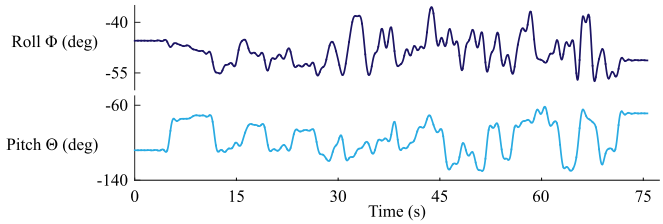

Fig. 7. These tilt angle signals were computed using the accelerometer signals of the right tool.

To compute the angular tilt (roll and pitch) of both tools and the camera, we first downsampled the associated three-axis acceleration signal to 100 Hz using the decimate function in MATLAB, which guards against aliasing, and we then low-pass filtered it using an eighth-order Butterworth infinite impulse response (IIR) filter with a 1-Hz cutoff frequency. These filtering steps largely eliminate contact vibrations and translational accelerations, leaving only a measure of the downward force of gravity. The sensor's roll (ϕ, rotation around the shaft) and pitch (θ, shaft angle relative to the horizontal) were computed according to the following equations:

$$\phi = \tan^{-1}\left(\frac{a_{\mathrm{fy}}}{a_{\mathrm{fz}}}\right) \tag{1}$$

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

where a_{fx}, a_{fy}, and a_{fz} are the downsampled and filtered x-axis, y-axis, and z-axis components of the acceleration vector. To account for potential nonunique solutions originating from the periodicity of the tangent function, we used MATLAB's two-argument inverse tangent function atan2 for all calculations. The resulting roll and pitch signals were then filtered using the same Butterworth IIR low-pass filter for additional smoothing. Fig. 7 shows an example of the roll and pitch signals of the right tool from one representative peg transfer trial. Note that we could not estimate the shaft's yaw angle because the measured direction of the gravity vector does not depend on this angle. Finally, we calculated the roll and pitch angle rates (velocities) over time as the difference in the respective orientation at pairs of successive time steps.

To process the acceleration data, the three-axis signals from each accelerometer were first mapped onto a single axis using the DFT321 algorithm [41]; the new signal preserves the spectral and temporal properties of the three-axis signals. The combined accelerometer signals were then segmented into

low-, medium-, and high-frequency bands using a Butterworth IIR low-pass filter with a 30-Hz stopband frequency, a 20-Hz passband frequency, a 65-dB stopband attentuation, and a 0.5-dB passband ripple; a Chebyshev type 1 eighth-order IIR bandpass filter with passband frequencies of 20 and 100 Hz and 0.1-dB passband ripple; and a Butterworth IIR high-pass filter with a 90-Hz stopband frequency, a 100-Hz passband frequency, a 65-dB stopband attenuation, and a 0.5-dB passband ripple.

1) Time Features: For each trial, we recorded the time elapsed from when the participant pressed the pedal to start data recording to the moment they pressed the pedal again to stop data recording. While some participants started the task immediately after the pedal press, many participants spent a few seconds adjusting the robot before beginning the task. Likewise, there were often a few seconds between when the participant finished the task and when he or she pressed the pedal to stop recording. Therefore, we also calculated the active trial completion time, defined as the difference between the last and first time the task board was touched (using a threshold of 0.25 N on the force magnitude). In addition, we calculated the square root and log of both the total completion time and the active completion time to account for the fact that relationships between skill and time may be nonlinear. The total number of time features was six.

2) Descriptive Features: For each trial, we computed the eight values of mean, standard deviation, minimum, maximum, range, RMS, TSS, and time integral of each of the following signals:

1) force in x-, y-, and z-directions (3 signals, 24 features);
2) force vector magnitude (one signal, eight features);
3) right tool, left tool, and camera roll and pitch angles (six signals, 48 features);
4) right tool, left tool, and camera roll and pitch angular velocities (six signals, 48 features);
5) right tool, left tool, and camera acceleration in each frequency band[1] (nine signals, 72 features);
6) product of right and left tool acceleration in each frequency band[1] (three signals, 24 features);
7) product of force magnitude and right tool acceleration in each frequency band[1] (three signals, 24 features);
8) product of force magnitude and left tool acceleration in each frequency band[1] (three signals, 24 features).

The product signals highlight when two items contact one another. The total number of descriptive features was 272.

IV. Machine Learning Algorithm Development

The final step in developing an automatic assessment tool for rating surgical skill is to train machine learning models to recognize patterns between the features and the ratings. Because GEARS scores are ordinal in nature, they can be treated as categories or as real-valued scores, suggesting both classification and regression approaches. This section explains how we trained our regression-based learners and classification-based learners to predict GEARS scores. For both, we created a testing set by reserving the trials from four participants who were randomly selected from the four groups of participants representing different reported familiarity levels with the da Vinci system. We then used the remaining 33 participants as a training set. Given that all participants completed the same number of trials except for one expert, this training/test split is close to the 90%/10% training/test split common in the machine learning literature. We trained and tested each learning model five times on a different training/test split to help account for the uneven distribution of ratings in our data. Since the GEARS ratings are integer values from 1 to 5, the rounded average rating between the three human raters was used for training.

A. Regression Learner

Before training our regression-based learner, we performed feature selection on our set of 278 features using stepwise regression, a search method that iteratively adds features to an initially empty model until it finds the model with the local minimum error, using an L_1 penalty. Feature selection was performed separately for each of the five GEARS domains. We then used leave-one-out cross-validation for model tuning and selection. Rather than using one specific regression technique, we trained an ensemble learner, which was composed of support vector regression [42], elastic net regression [43], regression trees [44], and K nearest neighbors [45]. The final model predictions are then an average of the predictions from each model. A separate ensemble learner was trained for each of the five GEARS domains. All regression learners were computed in MATLAB using the LIBSVM library [46], the Glmnet library [47], and the Statistics and Machine Learning Toolbox.

B. Classification Learner

We used a random forest classification learner [48], building a separate classifier for each of the five GEARS domains. Each classifier had 500 trees with a minimum leaf size of 25 for the Depth Perception, Bimanual Dexterity, and Force Sensitivity GEARS domains, and a minimum leaf size of 15 for the Efficiency and Robotic Control GEARS domains. The minimum leaf size was determined as the number of leaves that produced the smallest out-of-bag error, and for which additional leaves did not lower that error. The classification learner was built using the *TreeBagger* function in MATLAB's Statistics and Machine Learning Toolbox. We set the prior probabilities of each class to "empirical," which determined class probabilities based on the frequencies of classes in our training data. We set the cost matrix to be the squared error between ratings.

Because the Random Forest learner uses a bagging method to randomly sample the set of features for every tree built, feature importance is determined during model training using the out-of-bag error. We saved the importance of each feature and used only the 30 top features in each domain's final model.

V. Results

Both the regression learners and the classification learners produced meaningful automatic ratings on all five tested GEARS domains. The training and testing were performed on

[1]To avoid values near zero, the absolute value of the acceleration signal was used to calculate the mean of these signals.

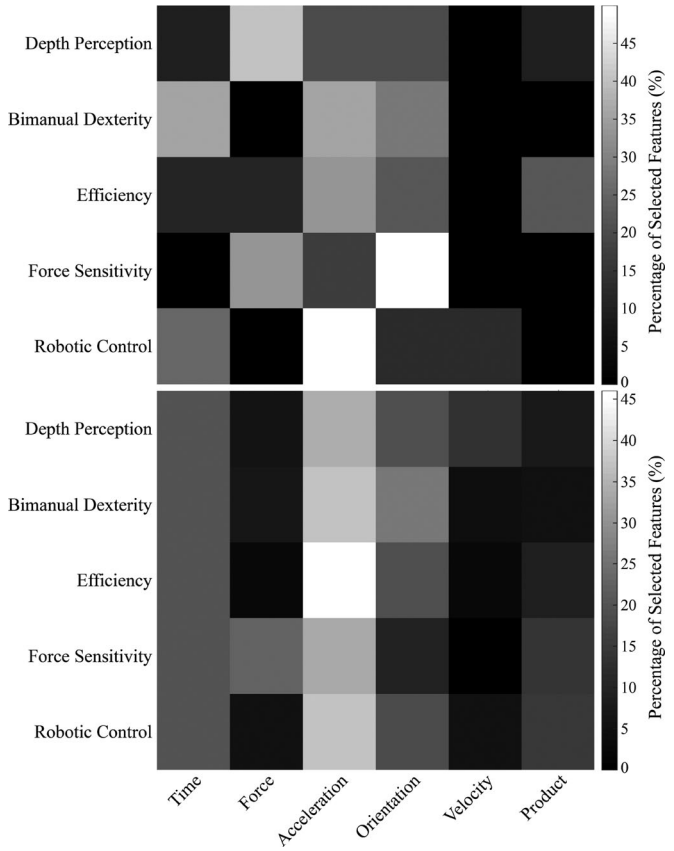

Fig. 8. Percentage of each type of feature (columns) that was important for each GEARS domain (rows) for the regression learner (top) and the classification learner (bottom).

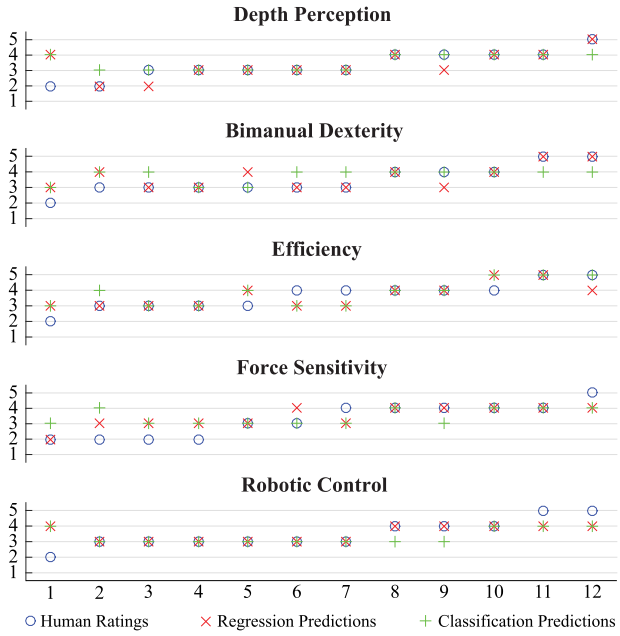

Fig. 9. Regression and classification learner predictions for all five GEARS domains for a representative testing set. Blue circles are the actual ratings from the human raters. Red ×'s are the average rounded predictions of the regression learner. Green +'s are the predictions of the Classification learner. The predictions are plotted for every trial in the testing set (1–12) in terms of increasing score (1–5).

a Mac OS X computer with a 2.8-GHz Intel i5 processor and 16 GB of RAM. On this machine, calculating the features for all 110 trials took approximately 18 min, training the five regression learners took approximately 22 s for all five GEARS domains, and training the five classification learners took approximately 33 min. Calculating all features and five ratings for a single task took approximately 11 s for regression learners and 18 s for the classification learners. The most important features for each GEARS domain differed between the two learners, as shown in Fig. 8.

All results reported below were obtained through analysis of the reserved testing sets, which were never seen during training. We evaluate the performance of each learner by 1) the accuracy with which it predicts the GEARS ratings given by our raters and 2) the interrater reliability between the automatic GEARS ratings and the GEARS ratings produced by our raters. We also analyze the performance of the learners after they were retrained without any force-based features, as the force sensor cannot be used during *in vivo* practice; a complete list of these findings can be found in Section IV-A of the online supplemental material.

A. Prediction Accuracy

We used the developed regression and classification learners to predict the five GEARS ratings for all trials by the four participants in each of the five reserved testing sets. Since the GEARS ratings are integer values from 1 to 5, we rounded the prediction from our regression learner, and we compared the predictions from both learners to the rounded average rating between the three human raters. Fig. 9 shows the resulting predictions for one of the reserved test sets.

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.

1) Exact Accuracy: On average, our classification learner and our regression learner were moderately accurate at predicting the exact ratings produced by the human raters. Table III shows the overall (mean ± standard deviation) accuracy for predicting the exact rating in each of the five GEARS domains for both learners across the five test sets. The accuracy for both the regression and classification learner was above 50% for all five GEARS domains, and the regression learner had a higher exact accuracy than the classification learner for all domains except Depth Perception and Robotic Control. The highest accuracy was for the classification learner in the Robotic Control domain (75.0%). The lowest accuracy was for the classification learner in the Force Sensitivity domain (51.7%).

TABLE IV
PRECISION AND RECALL FOR REGRESSION AND CLASSIFICATION

Precision							
	GEARS Domain		All Ratings	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$\left(\frac{tp}{tp+fp}\right)$	Depth Perception	(Regression)	0.47 ± 0.37	0.30 ± 0.45	0.59 ± 0.15	0.70 ± 0.15	0.25 ± 0.50
	Bimanual Dexterity	(Regression)	0.53 ± 0.38	0	0.58 ± 0.12	0.65 ± 0.18	0.80 ± 0.45
	Efficiency	(Regression)	0.64 ± 0.36	0	0.73 ± 0.25	0.66 ± 0.20	0.90 ± 0.22
	Force Sensitivity	(Regression)	0.43 ± 0.39	0.40 ± 0.55	0.44 ± 0.20	0.77 ± 0.08	0
	Robotic Control	(Regression)	0.47 ± 0.40	0.33 ± 0.58	0.80 ± 0.15	0.61 ± 0.22	0
	Depth Perception	(Classification)	0.40 ± 0.40	0	0.75 ± 0.25	0.69 ± 0.16	0
	Bimanual Dexterity	(Classification)	0.28 ± 0.30	0	0.52 ± 0.19	0.55 ± 0.20	0
	Efficiency	(Classification)	0.50 ± 0.38	0	0.60 ± 0.37	0.53 ± 0.14	0.87 ± 0.18
	Force Sensitivity	(Classification)	0.26 ± 0.28	0	0.37 ± 0.12	0.62 ± 0.09	0
	Robotic Control	(Classification)	0.44 ± 0.43	0	0.89 ± 0.15	0.62 ± 0.27	0
Recall							
	GEARS Domain		All Ratings	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$\left(\frac{tp}{tp+fn}\right)$	Depth Perception	(Regression)	0.56 ± 0.38	0.30 ± 0.45	0.76 ± 0.15	0.74 ± 0.10	0.33 ± 0.58
	Bimanual Dexterity	(Regression)	0.55 ± 0.38	0	0.68 ± 0.21	0.79 ± 0.13	0.60 ± 0.43
	Efficiency	(Regression)	0.65 ± 0.36	0	0.86 ± 0.13	0.80 ± 0.28	0.68 ± 0.21
	Force Sensitivity	(Regression)	0.45 ± 0.41	0.10 ± 0.14	0.71 ± 0.21	0.89 ± 0.10	0
	Robotic Control	(Regression)	0.55 ± 0.44	0.33 ± 0.58	0.92 ± 0.12	0.75 ± 0.19	0
	Depth Perception	(Classification)	0.48 ± 0.47	0	0.74 ± 0.20	1.00 ± 0.00	0
	Bimanual Dexterity	(Classification)	0.38 ± 0.41	0	0.62 ± 0.32	0.83 ± 0.13	0
	Efficiency	(Classification)	0.54 ± 0.33	0	0.45 ± 0.08	0.72 ± 0.23	0.78 ± 0.22
	Force Sensitivity	(Classification)	0.37 ± 0.40	0	0.61 ± 0.24	0.81 ± 0.20	0
	Robotic Control	(Classification)	0.51 ± 0.47	0	0.92 ± 0.12	0.82 ± 0.29	0

tp is the true positive result, fp is the false positive result, and fn is the false negative result. Values shown are mean $\pm$ standard deviation across the five testing sets for all ratings. Note that each value reported here is the mean of the individual scores, so the mathematical relationship between precision and recall does not necessarily follow the equations given. Note also that the training set contained significantly fewer examples of the ratings two and five than three and four. Furthermore, none of the five testing sets contained a rating of one, so the associated precision and recall scores have not been computed.

The exact accuracy for each learner in each domain was above the 20% expected for pure guessing with a uniform prior or the ~40% expected for always guessing the most common rating.

When the learners were retrained without the force-based features, there were only marginal differences ($\leq \pm 5\%$) in the average exact accuracy for both learners for all domains except for the regression learner in the Force Sensitivity domain, which decreased by 16.6%.

2) Prediction and Recall: In addition to the exact accuracy for the regression and classification learners, we also assessed the precision and recall score. Table IV reports these scores for the regression and classification learners averaged over all GEARS ratings (2–5, as there were no 1's in the test sets) as well as for each individual rating, averaged over all five test sets for each GEARS domain. When averaged together for all ratings, both learners produced precision scores above 0.2, indicating that performance was better than chance. In addition, the regression learner produced a higher precision and recall score than the classification learner for every GEARS domain. For the regression learner, in terms of individual ratings, the highest precision score was for Efficiency for a rating of five (0.90), and the highest recall score was for Robotic Control for a rating of three (0.92). For the classification learner, the highest precision score was for Robotic Control for a rating of three (0.89), and the highest recall score was for Depth Perception for a rating of four (1.0). Note that for all domains, the training set contained significantly fewer positive examples of ratings of two and five than three and four.

When the learners were retrained without the force-based features, there were only marginal changes ($\leq \pm 0.1$) in the precision and recall scores for both the learners when averaged together for all ratings, except for the regression learner in the Force Sensitivity domain, which decreased by 0.16 for precision.

B. Interrater Reliability of Learners

While the prediction accuracy explains how well each learner replicates the labels produced by the human raters, our ultimate goal is to use these machine learning techniques independent of a human rater. We, therefore, need to evaluate how well each learner would perform as an independent rater. As with our human raters (see Section III-A), we evaluated the interrater reliability of our learners using the two-way random-effect ICC. Table V shows the range of ICC values from our five testing sets for each of the individual GEARS domains, as well as the overall summed GEARS score. The ICC values reported in Table V quantify the reliability of the ratings based on the average ratings of all three raters and the respective learner. The median and upper range ICC values for both learners are in "excellent" agreement (ICC ≥ 0.75) with the human raters for all five GEARS domains and overall. The lower ICC values are all in "good" agreement (ICC ≥ 0.6) or better with the human raters.

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)

When the force-based features were removed, the ICC values changed only marginally ($\leq \pm$ 0.05). The median and upper range ICC values for both learners are still in "excellent" agreement for all five GEARS domains and overall. The lower range ICC values are also still in "good" agreement or better with the human raters.

VI. DISCUSSION

The presented results demonstrate that skill at robotic peg transfer can be assessed using only task completion time and the signals from accelerometers on the two robotic tools and camera and a force sensor under the task board. GEARS ratings were automatically predicted by machine learning algorithms that were trained with data labeled by human graders. Both regression and classification approaches were successful at predicting the human graders' ratings, but to slightly different degrees.

The exact accuracies for both learners was significantly above the performance expected for either random guessing or guessing the most common rating. In addition, the regression learner's average precision was above 0.4 for all ratings in all domains, while the classification learner's average precision was above 0.4 for all domains except Bimanual Dexterity and Force Sensitivity. Although many of the mean precision values for all ratings are lower, the precision values for ratings of three and four are above 0.5 for all domains except Force Sensitivity. The number of positive examples of a rating in the training set impacts the precision of the machine learning algorithm; as shown in Table II, most of the ratings for every domain were threes and fours. Looking at ratings of two and five, it appears classification is more negatively impacted by the low number of positive examples than regression. This trend largely holds true for the recall results, where the majority of the scores were greater than 0.6 for a rating of three or four.

The regression learner slightly outperformed the classification learner with few exceptions. This distinction most likely stems from the different algorithmic approaches of regression and classification. Regression treats the GEARS scores of 1–5 as real numbers with a sequentially ordered relationship. For classification, the GEARS scores of 1–5 are treated as separate categories with no underlying relationship. Apparently the relationship between the physical interaction data (tool and camera accelerations, contact forces, and completion time) and the rated skill of the participants can be described more accurately with a linear function than with a categorical algorithm.

While the prediction accuracy findings demonstrate the technical efficacy of the proposed machine learning approach, the interrater reliability findings highlight the potential impact this approach could have on the field of simulation-based surgical robotic training. The ICC values in Table V calculated across testing sets show that both the regression and classification learners are at a minimum in "good" reliability with the human raters, and at a maximum in "excellent" reliability, suggesting that both can be treated as independent evaluators. The median ICC value for both learners was above 0.75 for every GEARS domain and overall with regression performing slightly better than classification across the board. The minimum ICC values less than 0.75 for the Bimanual Dexterity, Force Sensitivity, and Robotic Control domains have more to do with the poor agreement of the human raters (minimum human ICC ratings of 0.60, 0.56, and 0.43, respectively) than with the machine learning techniques. Even though we attempted to keep them calibrated, these testing sets represent a few specific instances where the raters were not in good calibration with one another.

Given these findings, the regression-based approach appears to be more suited than classification for learning peg transfer skill ratings: it achieved better accuracy, precision, and recall; produced ratings that were reliable with those of the expert raters; took less time to train; and needs less time to make a prediction. The longer training and prediction times for the classification learner come from the higher computational load of Random Forests. In training these learners, the employed combination of 500 decision trees and at least 15 leaves produced the lowest out-of-bag errors, but other combinations may be possible. While fewer trees or leaves would reduce the computational cost, accuracy would likely also decrease. The regression learner, on the other hand, was trying to fit a linear model to the data, which is less computationally demanding.

Although both learners produced good prediction accuracies and high ICCs, certain domains performed better than others. Each GEARS domain covers a different aspect of surgical skill, so we expect that the human raters based each of their ratings on a particular subset of the behaviors visible in the video. Some of the features that we calculate from the recorded data, such as completion time, likely match up well with cues to which the raters attended, while other important visual cues may not have been captured sufficiently by our feature set.

It is interesting that the accuracy results were lowest for the Force Sensitivity domain. This finding may indicate that different features are needed to accurately capture the visual cues used by the expert raters. At the same time, however, it may allude to a potential confound of using stiff training tasks for which visual estimation of force is difficult. There may not be a universal visual cue that directly relates to Force Sensitivity for this task, so different raters may have relied on different cues. This hypothesis is supported by the fact that the human Force Sensitivity ratings varied more than those in other domains, achieving a reliability that was slightly less than excellent (0.74, see Table II). Although this large variability negatively affected the prediction accuracy of the learners, the ratings they produced

were still reasonably consistent with those of the human raters, earning a median ICC that was above 0.75. This discovery opens up the question of whether giving a human rater access to the same data recorded by the STB might change the rating he or she assigns to a trial, particularly for skill domains that are more difficult to rate visually.

Taken together, our findings have the potential to significantly impact simulation-based training for RMIS. RMIS is still young compared to other surgical approaches, but the number and type of RMIS procedures are growing in many surgical specialties [4]–[6]. Therefore, many novice surgeons will need to become proficient in the psychomotor skills needed to safely operate current and future RMIS platforms. While VR training will surely continue to play a role in skill development, training on the clinical robot will remain the gold standard because it provides the trainee with a real robotic experience, complete with robot dynamics and physical interactions between the instruments and the training task. In addition, with the advent of a standardized training curriculum, performance on the clinical robot will most likely be used for benchmarking purposes in the same manner in which real laparoscopic tools are used for the FLS [38].

A downside of clinical robotic training and evaluation, however, has been that expert human raters are needed to assess the trainee's skill level. This evaluation and assessment process can unfortunately be inefficient, not providing the trainee with immediate feedback. With the proposed approach, expert evaluators are needed only to provide ground-truth skill ratings for algorithm development. Once trained, however, these algorithms can automatically rate skill on their own, using a validated assessment tool, much faster than a human rater. As an illustration, it took our motivated expert surgical raters six months to rate 110 peg transfer trials (among their other clinical, research, and training responsibilities). Our regression learner could theoretically calculate the features and rate the same number of trials in approximately 20 min, and it would take our classification learner about 33 min. Thus, the proposed approach drastically reduces the time needed to obtain ratings. It can, therefore, provide trainees with real-time structured feedback, eliminating the need for expert surgical raters to assess basic psychomotor skill development.

Another benefit of the proposed approach is that it accounts for the actual physical interactions between the surgical robot and the training task, regardless of whether those interactions take place in the camera's field of view. This approach is, therefore, robust to any master–slave misalignment and compliance or mechanical wear in the robotic tools, and it does not depend on robot kinematics or any other knowledge of the surgical robot. Likewise, our STB system integrates on top of the existing surgical system and does not interfere with robot control and operation. Consequently, this approach can be applied to any surgical system, whether training or clinical, currently available or still in development. While we have demonstrated the efficacy of this approach for GEARS, it could be applied to other structured assessment tools. It could also be used to develop entirely new validated global assessment metrics, particularly metrics that account for the physical interactions that are hard to discern through vision, such as high-frequency tool accelerations and sustained contact forces.

While the proposed approach has demonstrated that the manner in which a surgeon brings the robot into contact with the training environment relates to surgical skill, many aspects of skill and technique are still grounded in the distinct motions produced through the robot by the surgeon. Indeed, many motion-based approaches have demonstrated efficacy in evaluating open and laparoscopic skill [20]–[23] as well as robotic surgical skill [24], [25]. It would then seem appropriate that these two approaches should be combined to more holistically evaluate skill. Motion and physical interaction data have already been combined to assess skill in endovascular catheterization [31], endoscopic sinus surgery [32], natural orifice translumenal endoscopic surgery [33], and laparoscopic surgery [34], [35]. They have even been used to assess laproscopic skill according to metrics comparable to those used in validated assessment tools [36], as we did in this study. These studies suggest the potential utility that exists for combining motion and physical interaction data for robotic surgical skill assessment.

The combination of motion and physical interaction data could also potentially contribute to skill assessment in the actual operating room. While the main results reported in this paper relied on data from a force sensor mounted beneath the task materials, we found only marginal reductions in performance for most domains when the signals from the force sensor were omitted, as they would need to be for assessment of clinical skills *in vivo*; unsurprisingly, the Force Sensitivity ratings were most affected. We believe the rest of our results were robust to this omission because instrument vibrations also somewhat capture the consequences of contact between the tools and the task materials. The STB accelerometers could be used during actual surgery if they were sterilized before attachment to the instruments and camera.

While the results presented in this paper demonstrate the efficacy of skill rating for robotic surgery based solely on physical interaction data, our claims have some limitations. First and foremost, we have demonstrated this approach only for the peg transfer task. Peg transfer is a widely accepted and validated training task for evaluating basic psychomotor skill. Still, it lacks the direct clinical relevance of a task like dissection or suturing. The use of soft tissue-like task materials would most likely improve the Force Sensitivity domain accuracy results because their visual deformation would give the raters a more universal indicator of applied force, potentially decreasing the variability in their Force Sensitivity ratings.

A second limitation is that we had only two expert surgical raters and one nonexpert rater. As mentioned previously, the experts in this study have significant experience both as robotic surgeons and as surgical graders, and the ICC metric was used to ensure reliability among all three raters. Still, there remains an element of subjectivity in the produced ratings. Likewise, although GEARS is a validated assessment tool, it was designed with the intent of evaluating surgical skill in the operating room. Here, we have utilized the universal aspects of the GEARS framework to assess skill in an inanimate *ex vivo* task. Given the good to excellent agreement between each learner and the

raters both with and without the force sensor, it seems very likely that the proposed approach would be effective in actual surgical training and evaluation, even in the operating room. This hypothesis needs to be tested in an actual training study for validation, where care should be taken to ensure the ratings produced by the learners are consistent with those produced by an human rater. Given that improper assessment with our approach carries with it the same consequences as for a human rater, the robustness of the current approach can be enhanced by training with a larger dataset and a more calibrated set of human raters.

A third limitation is that only 38 participants took part in the study, and as a result, not every skill level was equally represented. The effect of this limitation could be seen in the lower precision and recall scores for individual ratings of two and five. This dataset also precluded significant training on ratings of one for any domain. While even extremely novice participants scored mostly two and above for this task, neither learner is capable of accurately rating a trainee with extremely poor performance.

A fourth limitation is that all experiments were performed on a da Vinci Standard platform, which is no longer in clinical use in the U.S. Most of the experienced residents and fellows had prior experience with newer models of the da Vinci platform, and most experienced attending surgeons had to refamiliarize themselves with this older model. However, given that all sensors were external to the robot, updated algorithms could be developed for newer da Vinci models or other robotic platforms altogether.

Despite the aforementioned limitations, the approach demonstrated in this paper is the first automatic skill rating system that relies on physical interaction between the robot and the training task, and it achieved excellent predictions on unseen data. We, therefore, plan to conduct future studies to evaluate the presented methods on other training tasks, particularly ones that are more clinically relevant, such as suturing. We believe our methods will achieve good results even on tasks with softer materials because novice and experienced subjects will still differ in how they handle the tools and manipulate objects. Such investigations may prompt us to extend the methods reported here with new features and new machine learning approaches. These studies will ideally involve more participants from a broader range of skill levels, as well as more raters.

As a complementary approach, unsupervised machine learning techniques should be considered to allow skill assessment beyond the limited domains and 1–5 scale of GEARS and other similar assessments. Likewise, the current STB could be modified to work with newer robotic systems such as the da Vinci Xi, and it could include additional sensors on the instrument and camera shafts, such as gyroscopes and magnetometers. It would also probably prove worthwhile to integrate the approach proposed in this paper with motion-based skill assessment approaches. Finally, the utility of the ratings produced by this approach should be evaluated in a study aimed at helping novice trainees improve their skills with a robotic surgery system. We hypothesize that providing automatic GEARS scores after every trial will help trainees improve faster than they would with more common quantitative metrics (akin to the features themselves) or with no feedback at all.

VII. Conclusion

We created a system that can automatically evaluate a surgical trainee's skill at performing the common task of peg transfer with a robotic minimally invasive surgical system. This smart task board (STB) rates skill using the GEARS structured assessment tool, which involves 1–5 ratings in five domains. Its ratings are based on the manner in which the trainee brought the robot into physical contact with the training task materials, which is measured using force, acceleration, and time sensors that are external to the robot. GEARS ratings are then predicted using custom regression-based and classification-based machine learning algorithms whose feature set is calculated from signals produced by the external sensors. Training of both the regression and classification learners was performed using peg transfer data from participants of various skill level and was labeled by two expert robotic surgeons and one nonexpert rater. Both approaches produced highly accurate and reliable GEARS predictions on unseen data even when the force-based features were removed. Regression, however, outperformed classification in terms of both prediction accuracy and computation time, making it the superior choice for this particular form of inanimate task training. This study is the first to demonstrate automatic skill assessment for RMIS via physical interaction information. It makes significant progress toward the goal of improved surgical training and evaluation by reducing the need for human raters to assess basic psychomotor skill development.

Acknowledgements

The authors would like to thank R. Beato (University of Pennsylvania) and A. Jarc (Intuitive Surgical) for their contributions to the project.

References

[1] J. L. Cameron, "William Stewart Halsted. our surgical heritage," *Ann. Surgery*, vol. 225, no. 5, pp. 445–458, 1997.
[2] C. B. Barden *et al.*, "Effects of limited work hours on surgical training," *J. Amer. Coll. Surgeons*, vol. 195, no. 4, pp. 531–538, 2002.
[3] J. Y. Lee *et al.*, "Validation study of a virtual reality robotic simulator—Role as an assessment tool?," *J. Urol.*, vol. 187, no. 3, pp. 998–1002, 2012.
[4] K. K. Badani *et al.*, "Evolution of robotic radical prostatectomy," *Cancer*, vol. 110, no. 9, pp. 1951–1958, 2007.
[5] A. L. Smith *et al.*, "Survey of obstetrics and gynecology residents' training and opinions on robotic surgery," *J. Robot. Surgery*, vol. 4, no. 1, pp. 23–27, 2010.
[6] S. Maeso *et al.*, "Efficacy of the da Vinci surgical system in abdominal surgery compared with that of laparoscopy: A systematic review and meta-analysis," *Ann. Surgery*, vol. 252, no. 2, pp. 254–262, 2010.
[7] G. H. Ballantyne, "Robotic surgery, telerobotic surgery, telepresence, and telementoring. Review of early clinical results," *Surgical Endoscopy Interventional Techn.*, vol. 16, no. 10, pp. 1389–402, 2002.
[8] G. Guthart and J. Salisbury, "The Intuitive telesurgery system: Overview and application," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2000, pp. 618–621.
[9] M. E. Hagen *et al.*, "Visual clues act as a substitute for haptic feedback in robotic surgery," *Surgical Endoscopy Interventional Techn.*, vol. 22, no. 6, pp. 1505–1508, 2008.
[10] R. Smith *et al.*, "Fundamentals of robotic surgery: A course of basic robotic surgery skills based upon a 14-society consensus template of outcomes measures and curriculum development," *Int. J. Med. Robot. Comput. Assisted Surgery*, vol. 10, no. 3, pp. 379–384, 2014.
[11] A. J. Hung *et al.*, "Comparative assessment of three standardized robotic surgery training methods," *BJU Int.*, vol. 112, no. 6, pp. 864–871, 2013.

[12] R. Korets *et al.*, "Validating the use of the Mimic dV-trainer for robotic surgery skill acquisition among urology residents," *The J. Urol.*, vol. 78, no. 6, pp. 1326–1330, 2011.
[13] I. Nisky *et al.*, "Uncontrolled manifold analysis of arm joint angle variability during robotic teleoperation and freehand movement of surgeons and novices," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 12, pp. 2869–2881, Dec. 2014.
[14] I. Nisky *et al.*, "Effects of robotic manipulators on movements of novices and surgeons," *Surgical Endoscopy Interventional Techn.*, vol. 28, no. 7, pp. 2145–2158, 2014.
[15] P. Culligan *et al.*, "Predictive validity of a training protocol using a robotic surgery simulator," *Female Pelvic Med. Reconstructive Surgery*, vol. 20, no. 1, pp. 48–51, 2014.
[16] A. I. Tergas *et al.*, "A pilot study of surgical training using a virtual robotic surgery simulator," *J. Soc. Laparoendoscopic Surgeons*, vol. 17, no. 2, pp. 219–26, 2013.
[17] A. J. Hung *et al.*, "Concurrent and predictive validation of a novel robotic surgery simulator: A prospective, randomized study," *J. Urology*, vol. 187, no. 2, pp. 630–637, 2012.
[18] C. Chen *et al.*, "Crowd-sourced assessment of technical skills: A novel method to evaluate surgical performance," *J. Surgical Res.*, vol. 187, no. 1, pp. 65–71, 2014.
[19] C. E. Reiley *et al.*, "Review of methods for objective surgical skill evaluation," *Surgical Endoscopy Interventional Techn.*, vol. 25, no. 2, pp. 356–366, 2011.
[20] V. Datta *et al.*, "The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model," *J. Amer. Coll. Surgeons*, vol. 193, no. 5, pp. 479–85, 2001.
[21] G. M. Fried *et al.*, "Proving the value of simulation in laparoscopic surgery," *Ann. Surgery*, vol. 240, no. 3, pp. 518–528, 2004.
[22] J. Rosen *et al.*, "Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills," *IEEE Trans. Biomed. Eng.*, vol. 48, no. 5, pp. 579–591, May 2001.
[23] C. Cao *et al.*, "Task and motion analysis in endoscopic surgery," in *Proc. IEEE Symp. Haptic Interfaces Virtual Environ. Teleoperator Syst.*, 1996, pp. 583–590.
[24] H. C. Lin *et al.*, "Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions," *Comput. Aided Surgery*, vol. 11, no. 5, pp. 220–230, 2006.
[25] N. Ahmidi *et al.*, "String motif-based description of tool motion for detecting skill and gestures in robotic surgery," in *Proc. Med. Image Comput. Comput.-Assisted Intervention Conf.*, 2013, pp. 26–33.
[26] L. H. Kim *et al.*, "Effects of master-slave tool misalignment in a teleoperated surgical robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 5364–5370.
[27] S. Speidel *et al.*, "Tracking of instruments in minimally invasive surgery for surgical skill analysis," in *Proc. 3rd Int. Workshop Med. Imag. Augmented Reality*, 2006, pp. 148–155.
[28] L. Zappella *et al.*, "Surgical gesture classification from video and kinematic data," *Med. Image Anal.*, vol. 17, no. 7, pp. 732–745, 2013.
[29] K. Bark *et al.*, "Surgical instrument vibrations are a construct-valid measure of technical skill in robotic peg transfer and suturing tasks," in *Proc. Hamlyn Symp. Med. Robot.*, 2012, pp. 50–51.
[30] E. D. Gomez *et al.*, "Objective assessment of robotic surgical skill using instrument contact vibrations," *Surgical Endoscopy Interventional Techn.*, vol. 30, pp. 1419–1431, 2015.
[31] H. Rafii-Tari *et al.*, "Towards automated surgical skill evaluation of endovascular catheterization tasks based on force and motion signatures," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2015, pp. 1789–1794.
[32] Y. Yamauchi *et al.*, "Surgical skill evaluation by force data for endoscopic sinus surgery training system," *Med. Image Comput. Comput.-Assisted Intervention*, vol. 2488, pp. 44–51, 2002.
[33] S. Dargar *et al.*, "Characterization of force and torque interactions during a simulated transgastric appendectomy procedure," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 890–899, Mar. 2015.
[34] H. Hwang *et al.*, "Correlating motor performance with surgical error in laparoscopic cholecystectomy," *Surgical Endoscopy Interventional Techn.*, vol. 20, no. 4, pp. 651–655, 2006.
[35] J. Rosen *et al.*, "Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 3, pp. 399–413, Mar. 2006.
[36] A. L. Trejos *et al.*, "Development of force-based metrics for skills assessment in minimally invasive surgery," *Surgical Endoscopy Interventional Techn.*, vol. 28, no. 7, pp. 2106–2119, Jul. 2014.
[37] A. C. Goh *et al.*, "Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills," *J. Urol.*, vol. 187, no. 1, pp. 247–252, 2012.
[38] Fundamentals of Laparoscopic Surgery, 2016. [Online]. Available: http://www.flsprogram.org/
[39] K. A. Hallgren, "Computing inter-rater reliability for observational data: An overview and tutorial," *Tuts. Quant. Methods Psychol.*, vol. 8, no. 1, pp. 23–34, 2012.
[40] M. Pedley, "Tilt sensing using a three-axis accelerometer," Freescale Semiconductor Appl. Notes, 2013, pp. 1–22.
[41] N. Landin *et al.*, "Dimensional reduction of high-frequency accelerations for haptic rendering," in *Haptics: Generating and Perceiving Tangible Sensations SE-12* (ser. Lecture Notes in Computer Science), vol. 6192, A. Kappers, *et al.*, Ed. SBerlin, Germany: Springer, 2010, pp. 79–86.
[42] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[43] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. B (Statist. Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.
[44] L. Breiman *et al.*, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.
[45] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.
[46] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.
[47] J. Friedman *et al.*, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010.
[48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

Jeremy D. Brown (S'11–M'14) received the Ph.D. degree in mechanical engineering from the University of Michigan, Ann Arbor, MI, USA, in 2014.

He is a Postdoctoral Research Fellow in the Department of Mechanical Engineering and Applied Mechanics and the Haptics Group in the GRASP Laboratory, University of Pennsylvania, Philadelphia, PA, USA. His research focuses on the interface between humans and robots with a specific focus on medical applications and haptic feedback.

Dr. Brown received several awards including the National Science Foundation Graduate Research Fellowship, the Best Student Paper Award from the IEEE Haptics Symposium in 2012, and the Penn Postdoctoral Fellowship for Academic Diversity.

Conor E. O'Brien received the B.S. degree in mechanical engineering and applied mechanics and the M.S. degree in robotics from the University of Pennsylvania, Philadelphia, PA, USA, in 2012 and 2015, respectively.

Sarah C. Leung received the M.S. degree in robotics from the University of Pennsylvania, Philadelphia, PA, USA, in 2014, with a background in mechanical and aerospace engineering from Cornell University, Ithaca, NY, USA.

She has been an Engineer at Boeing since 2010, previously working on CH-47 structural design in Philadelphia, and currently working on robotic systems for automated fiber placement at the Composite Fabrication Laboratory in Charleston, SC, USA.

Kristoffel R. Dumon completed his residency in general surgery at the University of Pennsylvania, Philadelphia, PA, USA, in 2006.

He is an Assistant Professor of surgery at the University of Pennsylvania. He specializes in minimally invasive and robotic surgery. His research interests include innovation in simulation with a focus on integrating new technology in simulation and its translation to the surgical clinical setting.

Prof. Dumon is a member of multiple professional societies and the Research Director of the Penn Medicine Clinical Simulation Center, an American College of Surgeons Level 1 Accredited Educational Institute.

David I. Lee completed his residency in urology at Thomas Jefferson University Hospital, Philadelphia, PA, USA, in 2001, followed by an endourology fellowship at the University of California Irvine Medical Center, Orange, CA, USA, where he was part of one of the pioneering experiences of robotic surgery. He is the Chief of Urology at the Penn Presbyterian Medical Center and an Assistant Professor of surgery and urology at the Perelman School of Medicine, University of Pennsylvania. He routinely performs more than 400 robotic cases per year and has completed more than 4500 robotic cases overall. He has published more than 100 articles, abstracts, and book chapters in the field of minimally invasive urologic surgery. He is the Director of the UPHS Robotic Training Center and the Penn robotic urologic surgery fellowship and a member of the Abramson Cancer Center.

Prof. Lee has been recognized in Best Doctors in America, Castle Connelly Top Doctors, and Top Docs in Philadelphia Magazine.

Katherine J. Kuchenbecker (S'04–M'06) received the B.S., M.S., and Ph.D. degrees in mechanical engineering from Stanford University, Stanford, CA, USA, in 2000, 2002, and 2006, respectively.

She completed a Postdoctoral Research Fellowship at The Johns Hopkins University, Baltimore, MD, USA, in 2006–2007. She is currently the Class of 1940 Bicentennial Endowed Term Chair and an Associate Professor in mechanical engineering and applied mechanics at the University of Pennsylvania, Philadelphia, PA, USA. Her research interests include the design and control of haptic interfaces and robotic systems. She directs the Penn Haptics Group, which is part of the General Robotics, Automation, Sensing, and Perception Laboratory.

Prof. Kuchenbecker received the 2009 National Science Foundation CAREER Award and the 2012 IEEE Robotics and Automation Society Academic Early Career Award. She is co-chairing the IEEE Haptics Symposium in 2016 and 2018. She is a co-chair of the IEEE Robotics and Automation Society Technical Committee on Haptics.