

Predicting surgical skill from the first N seconds of a task: value over task time using the isogony principle

Anna French¹  · Thomas S. Lendvay² · Robert M. Sweet³ · Timothy M. Kowalewski¹

Received: 9 February 2017 / Accepted: 3 May 2017 / Published online: 17 May 2017
© CARS 2017

Abstract

Purpose Most evaluations of surgical workflow or surgeon skill use simple, descriptive statistics (e.g., time) across whole procedures, thereby deemphasizing critical steps and potentially obscuring critical inefficiencies or skill deficiencies. In this work, we examine off-line, temporal clustering methods that chunk training procedures into clinically relevant surgical tasks or steps during robot-assisted surgery.

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 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 8 s 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. The accuracy of this classification comes within 90% of the classification

accuracy from observing the full trial within 10 s of task initiation on average.

Keywords Surgical skill evaluation · Computer aided decision · Tracking systems

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 (surges) required a surgeon's expertise. Investigation from [1] reports similar classification success but also used features very specific to septoplasty, and classified by seg-

✉ Anna French
afrench@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

menting 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 are 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?

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:

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

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.

Methods

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 were 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 30 Hz 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.

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 were recorded at 30 Hz, so $n = 30t$):

$$\mu({}^t\gamma_d) = \frac{1}{n(t)} \sum_{i=0}^{n(t)} \frac{v_d(i)}{k_d(i)^{1/3}}$$

$$\sigma({}^t\gamma_d) = \sqrt{\frac{1}{n(t)} \sum_{i=0}^{n(t)} \left(\frac{v_d(i)}{k_d(i)^{1/3}} - \mu({}^t\gamma_d) \right)^2}$$

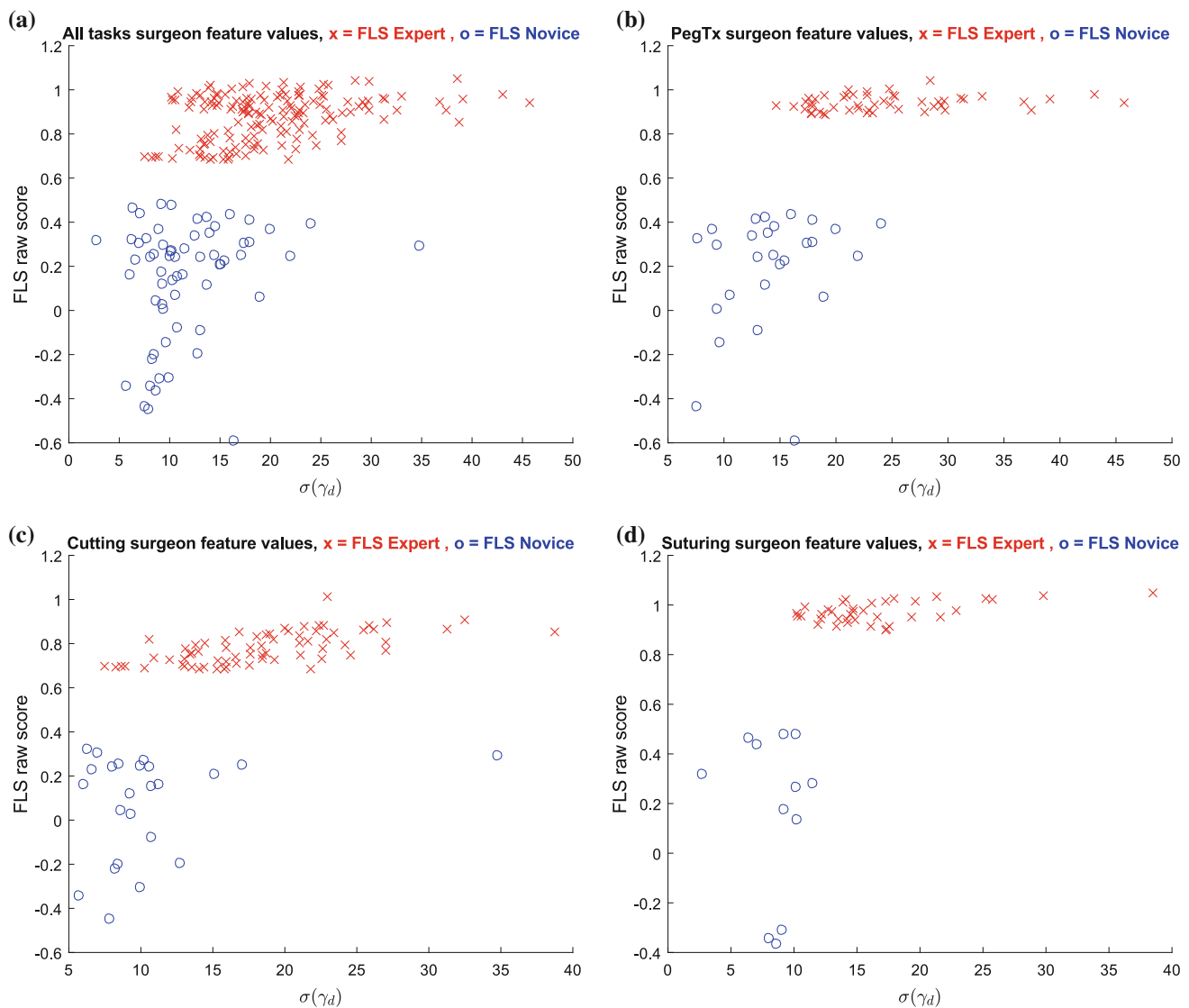


Fig. 1 FLS score versus $\sigma(\gamma_d)$ feature calculated over the full duration of each trial (dominant hand). Each point represents the feature value for one trial. Raw data displayed for all trials and each trial individually. **a** All trials, **b** peg transfer trials, **c** cutting trials, **d** suturing trials

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 30 s was chosen since all task times in the data set were greater than this amount. This created 30 different 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 was 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 train three different models using either tool path lengths (PL), economy of motion (EOM) and motion smoothness (MS) as features using SVM.

Experimental results

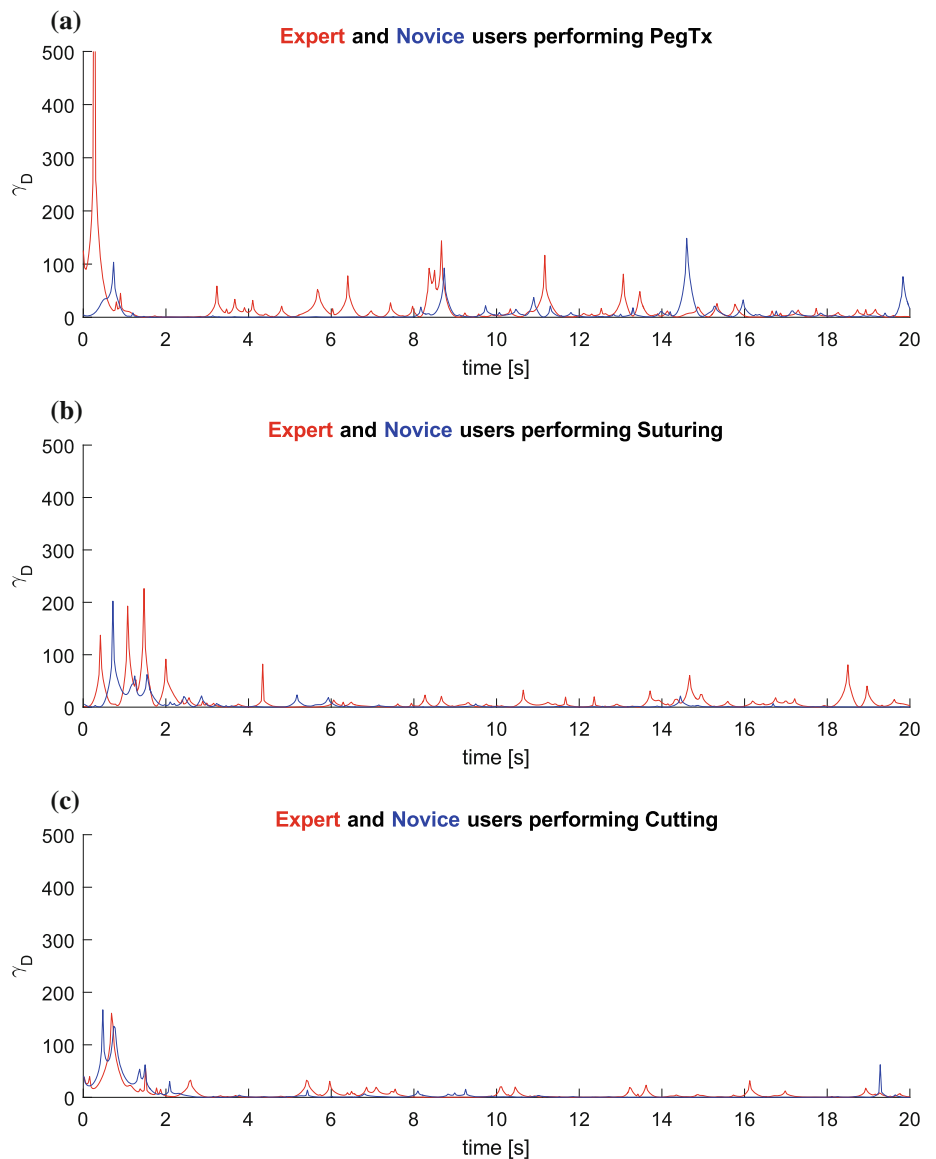
Two class classification: FLS expert versus FLS novice

Figure 1a–d shows 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-

Table 1 Median and standard deviation of the $\sigma(\gamma_d)$ and $\sigma(\gamma_{nd})$ over all subjects

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

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



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.

Figure 2a–c shows six example plots of the γ_D parameter calculated for each time step in the first 20s of a trial.

Figure 3 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

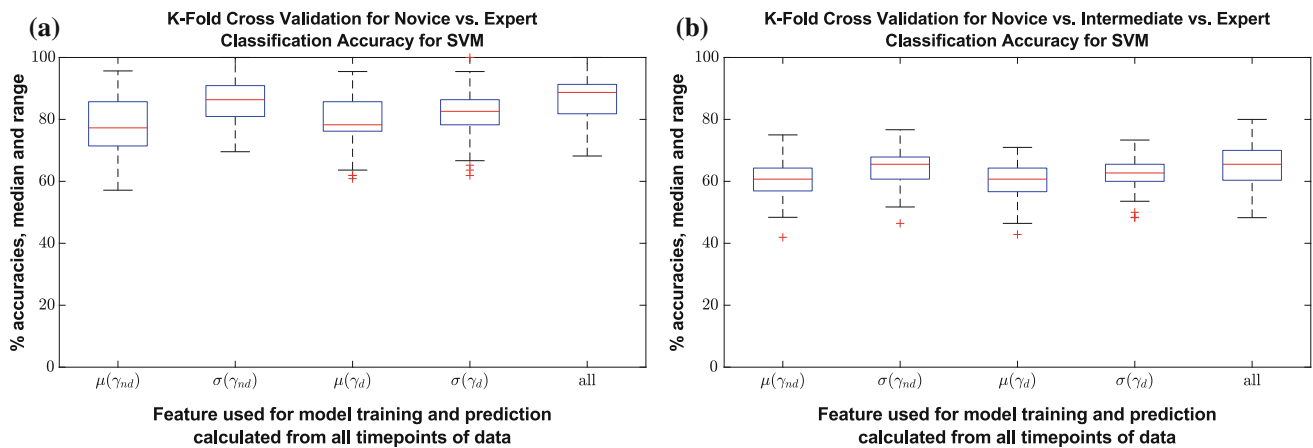


Fig. 3 K-fold leave-trial-out SVM classification task-blind training and task-blind testing, comparing strength of each feature. **a** 2-class classification, **b** 3-class classification

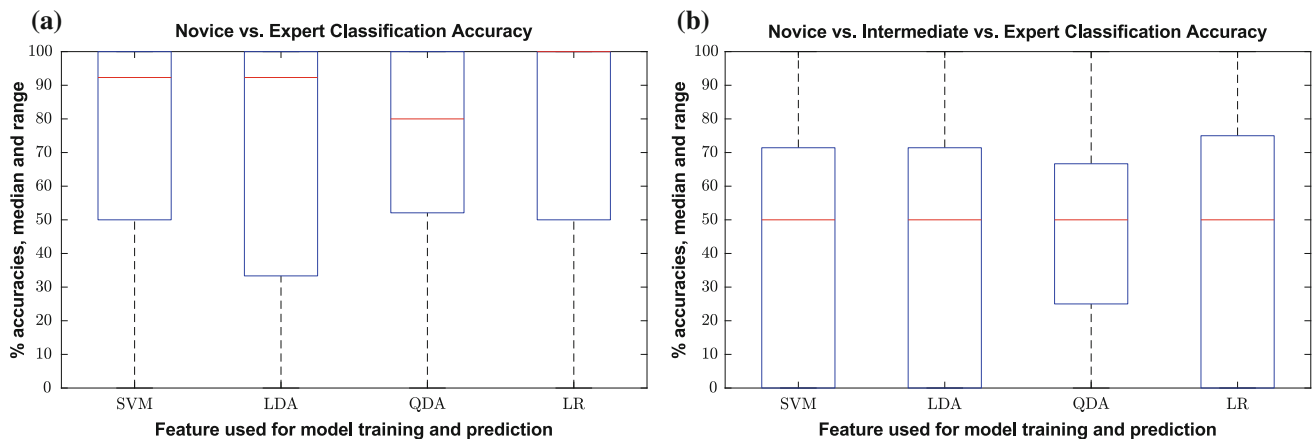


Fig. 4 LOUO classification task-blind training and task-blind testing using $\mu(\gamma_d)$, $\sigma(\gamma_d)$, $\mu(\gamma_{nd})$, and $\sigma(\gamma_{nd})$ features combined to train each model, comparing strength of machinery. **a** 2-class classification, **b** 3-class classification

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.

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.

Three class classification: FLS expert, FLS intermediate, FLS novice

Figure 3b uses the same k-fold method as Fig. 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 using all features. Figure 4b uses the same LOUO method as Fig. 4a, and it also predicts over all three classes. It shows a median agreement of 50%.

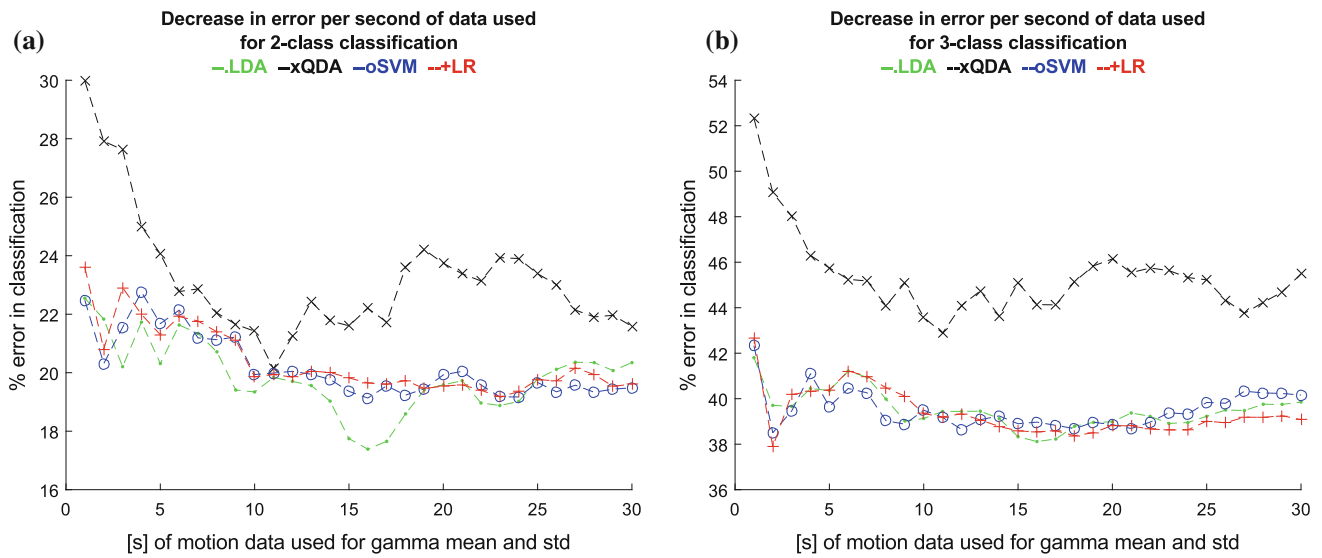


Fig. 5 Error rate per number of seconds used to calculate $\mu(t\gamma_d)$, $\sigma(t\gamma_d)$, $\mu(t\gamma_{nd})$, and $\sigma(t\gamma_{nd})$. **a** 2-class classification, **b** 3-class classification

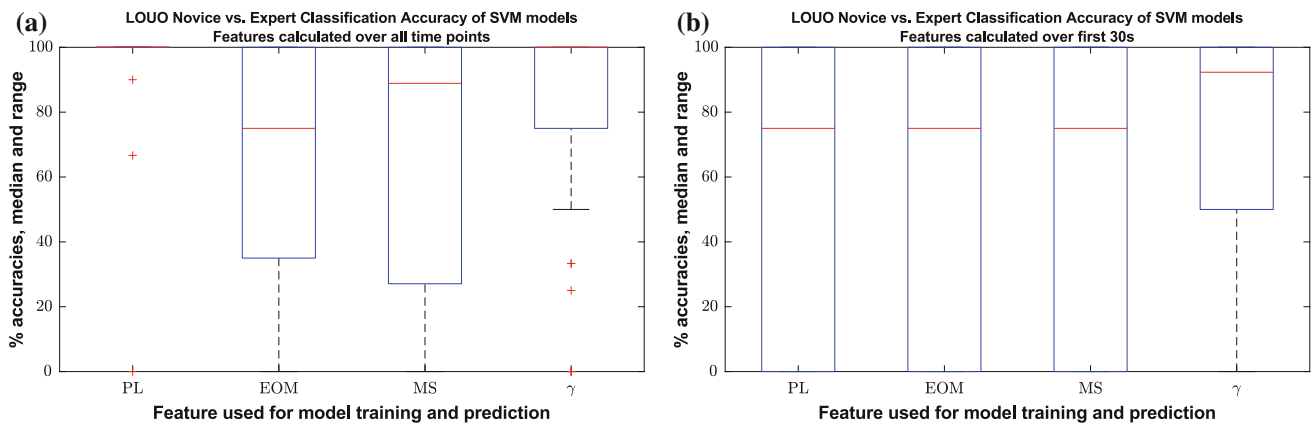


Fig. 6 LOUO validation for a 2-class classification using SVM for task-blind training and task-blind testing, comparing strength of each feature set. **a** Features calculated at end of task time, **b** features calculated at 30s point in task time

Table 2 2-class classification accuracy results per algorithm and per task

	Task-specific accuracy (μ) over all kN model accuracies in cross-validation, mean(SD)			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 (0.13)	86.1 (0.11)	83.0 (0.2)	83.7/85.7	2 s
QDA	87.6 (0.12)	86.9 (0.11)	66.0 (0.24)	82.5/82.6	4 s
SVM	85.2 (0.12)	90.2 (0.10)	82.6 (0.20)	87.2/88.7	8 s
LR	84.6 (0.12)	90.2 (0.10)	81.5 (0.21)	86.6/86.4	7 s

Models generated for this figure were trained task-blind, results at left reports how well the model classifies each task

Minimum time to classification

Figure 5a is 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 0 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,

Table 3 3-class classification accuracy results per algorithm

	Task-specific accuracy (μ) over all kN model accuracies in cross-validation, mean(SD)			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 (0.13)	63.0 (0.14)	66.0 (0.16)	61.6/61.7	2 s
QDA	66.6 (0.12)	59.2 (0.14)	44.3 (0.16)	58.9/58.6	4 s
SVM	62.2 (0.13)	67.0 (0.13)	65.5 (0.14)	65.1/65.5	3 s
LR	60.8 (0.12)	68.3 (0.13)	65.7 (0.15)	65.1/64.3	3 s

but is classifying between novice, intermediate and expert surgeons.

Validation with similar methods

Figure 6a and b compares the performance for 2-class classification of the γ parameters against previously validated aggregate task metrics as described in [3]. PL_d and PL_{nd} are the path length variables calculated for the dominant and non-dominant hand. The PL boxes represent 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 $\sigma(\gamma_{nd})$.

Conclusion

The results give support to our three initial hypotheses. The first hypothesis is supported by Fig. 1a–d 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 $\sigma(\gamma_d)$, $\mu(\gamma_d)$, $\sigma(\gamma_{nd})$, and $\mu(\gamma_{nd})$ 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 have insufficient N to capture human variability in surgery.

Third, Fig. 5a, b and Tables 2 and 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 Fig. 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, Fig. 6a and b shows that γ parameters outperform economy of motion and motion smoothness, but fall short of path length. When measured short of task completion (at 30s), Fig. 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 30 s 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 laparoscopy 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 30 s 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 s) 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 toward 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 s of a trial is potentially feasible and that isogony

provides some useful task-blind skill-classification information above simple task-time or FLS score.

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.

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 Helsinki declaration and its later amendments or comparable ethical standards.

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

References

- 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
- 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 Engl J Med* 369(15):1434–1442
- 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. *Surg Endosc* 21(11):2069–2075
- Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL (1998) Development of a model for training and evaluation of laparoscopic skills. *Am J Surg* 175(6):482–487
- Fried GM (2008) FLS assessment of competency using simulated laparoscopic tasks. *J Gastrointest Surg* 12(2):210–212. doi:10.1007/s11605-007-0355-0
- Kowalewski TM (2012) Real-time quantitative assessment of surgical skill. PhD thesis, University of Washington
- 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
- 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. *J Urol* 195(6):1859–1865
- Lacquaniti F, Terzuolo C, Viviani P (1983) The law relating the kinematic and figural aspects of drawing movements. *Acta Psychol* 54(1–3):115–130
- 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
- Peters JH, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, Hoffman K (2004) Development and Validation of a com-

- prehensive 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. *J Cataract Refract Surg* 35(11):1956–1960
 13. Shute VJ (2008) Focus on formative feedback. *Rev educ Res* 78(1):153–189
 14. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayed 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