Int J CARS (2015) 10:1435–1447
DOI 10.1007/s11548-015-1238-6

ORIGINAL ARTICLE

A study of crowdsourced segment-level surgical skill assessment using pairwise rankings

Anand Malpani[1] · S. Swaroop Vedula[1] · Chi Chiung Grace Chen[2] · Gregory D. Hager[1]

Received: 15 December 2014 / Accepted: 4 June 2015 / Published online: 30 June 2015

Abstract
Purpose Currently available methods for surgical skills assessment are either subjective or only provide global evaluations for the overall task. Such global evaluations do not inform trainees about where in the task they need to perform better. In this study, we investigated the reliability and validity of a framework to generate objective skill assessments for segments within a task, and compared assessments from our framework using crowdsourced segment ratings from surgically untrained individuals and expert surgeons against manually assigned global rating scores.
Methods Our framework includes (1) a binary classifier trained to generate *preferences* for pairs of task segments (i.e., given a pair of segments, specification of which one was performed better), (2) computing segment-level percentile scores based on the preferences, and (3) predicting task-level scores using the segment-level scores. We conducted a crowdsourcing user study to obtain manual preferences for segments within a suturing and knot-tying task from a crowd of surgically untrained individuals and a group of experts. We analyzed the inter-rater reliability of preferences obtained from the crowd and experts, and investigated the validity of task-level scores obtained using our framework. In addition, we compared accuracy of the crowd and expert preference classifiers, as well as the segment- and task-level scores obtained from the classifiers.
Results We observed moderate inter-rater reliability within the crowd (Fleiss' kappa, $\kappa = 0.41$) and experts ($\kappa = 0.55$). For both the crowd and experts, the accuracy of an automated classifier trained using all the task segments was above par as compared to the inter-rater agreement [crowd classifier 85 % (SE 2 %), expert classifier 89 % (SE 3 %)]. We predicted the overall global rating scores (GRS) for the task with a root-mean-squared error that was lower than one standard deviation of the ground-truth GRS. We observed a high correlation between segment-level scores ($\rho \geq 0.86$) obtained using the crowd and expert preference classifiers. The task-level scores obtained using the crowd and expert preference classifier were also highly correlated with each other ($\rho \geq 0.84$), and statistically equivalent within a margin of two points (for a score ranging from 6 to 30). Our analyses, however, did not demonstrate statistical significance in equivalence of accuracy between the crowd and expert classifiers within a 10 % margin.
Conclusions Our framework implemented using crowdsourced pairwise comparisons leads to valid objective surgical skill assessment for segments within a task, and for the task overall. Crowdsourcing yields reliable pairwise comparisons of skill for segments within a task with high efficiency. Our framework may be deployed within surgical training programs for objective, automated, and standardized evaluation of technical skills.

Keywords Robotic surgery · Training · Skill assessment · Feedback · Task flow · Crowdsourcing · Pairwise comparisons · Activity segments · Task decomposition

✉ Anand Malpani
anandmalpani@jhu.edu

Gregory D. Hager
hager@cs.jhu.edu

[1] Johns Hopkins University, 3400 N Charles St, Hackerman Hall Room 200, Baltimore, MD, USA

[2] Johns Hopkins Bayview Medical Center, 301 Building, 301 Mason Lord Drive, Room 3200, Baltimore, MD, USA

Introduction

Objective assessment of surgical technical skills and competence is an integral part of graduate surgical training curricula. Surgeons must be technically competent to provide safe and effective patient care. Inferior technical skills are associated with a higher risk of postoperative complications, including readmission, reoperation, and death [4]. Surgical trainees acquire technical skills through observation in the operating room and deliberate practice. Traditionally, faculty surgeons imparted technical skills to trainees in the operating room, and assessed trainees' technical skills using subjective, non-standardized measures. Recent policy set forth by the American Council of Graduate Medical Education (ACGME), the governing body for graduate medical training in the United States, mandated that trainees' competence (including technical competence) be determined using objective measures [28].

Lack of efficiently computed, reliable, and valid objective measures is a major limitation for academic surgical training programs in implementing ACGME's policy mandate. Currently available methods for surgical technical skills assessment rely upon the subjective opinion of faculty surgeons. For example, structured skills assessment tools such as the Objective Structured Assessment of Technical Skills (OSATS), Global Operative Assessment of Laparoscopic Skills (GOALS), and Global Evaluative Assessment of Robotic Skills (GEARS) require manual evaluation by the supervising surgeon [14,22,30]. Thus, use of these tools within surgical skills training curricula is limited by availability of faculty time.

Several methods have been developed for objective assessment of technical skills to supplement or substitute subjective manual evaluations. These objective methods use data captured while surgeons perform the operation. Some simple methods include computing measures of time and motion efficiency [9,10,16,17]. Other methods involve modeling surgical tool motion data or video images of surgical task performance, and derive objective measures of skill based on the models. Previous works have explored various approaches for developing such models, including graphical models [23–25,29], and linear dynamical systems [15,31] using tool motion or video data, or both [26,31].

In this paper, we explore two key ideas. First, OSATS, GOALS, and GEARS, and other comparable alternatives (based on tool motion or video data) for skill assessment only provide a global evaluation of surgeons' skills. Such global, task-level assessments do not inform trainees about where in the task they need to perform better in order to operate like an expert. In contrast, we would expect that skill assessment at the level of meaningful semantic segments may be more effective for skill acquisition in trainees. However, segment-level skill assessment is challenging because it requires significant manual resources to both segment and assess skill at a finer level of granularity than existing task-level methods. Furthermore, no existing reliable and valid tools exist to do so.

Second, we note that crowdsourcing has been utilized in the medical imaging domain to train image classifiers [19] as well as to generate reference correspondence regions in endoscopic images [20] with success. Similarly, a prior study has shown that crowdsourcing is an effective means of generating absolute surgical skill assessment based on GEARS [5]. However, it has proven difficult to perform absolute assessment of segment-level skill. Pairwise comparisons have been shown to yield valid assessments when absolute assessment is difficult—examples include assessing disease severity, movie recommendations, and information retrieval [11,13,18]. Thus, pairwise comparisons performed by a crowd may provide efficient, reliable, and valid solutions for objective assessment of segment-level surgical technical skills.

In a previous pilot study, using a limited sample, we demonstrated that crowdsourcing can yield reliable and valid pairwise comparison of surgical skill at the segment-level [21]. In this paper, we extend our analysis with a larger sample size, and also explore the computation and validation of global rating scores using ranking-based methods.

In summary, our goals in this paper are: (1) to establish reliability and validity of a framework to objectively assess surgical skill using pairwise comparisons of task segments, and (2) to compare assessments obtained from our framework using pairwise comparisons from two sources—a surgically untrained crowd and a group of expert surgeons. The remainder of this paper is structured as follows: we describe our framework for objective surgical skill assessment using pairwise comparisons of task segments in the "Methods" section, the user study and experimental setup for validating our framework in the "Experiments" section, results from our analyses in the "Results" section, discussion on the results and limitations of our study in the "Discussion" section, and our conclusions in the final section.

Methods

Our skill assessment framework consists of three components as shown in Fig. 1. The first component is an automated classifier to assign skill-based preferences in pairwise comparisons of task segments. We then use this classifier to compute percentile scores for task segments as an objective measure of segment-level skill. Finally, we compute an OSATS-like score for the overall task using the segment-level percentile skill scores.

Preference classifier

The first component in our framework is a binary classifier that selects the better-performed task segment from a given

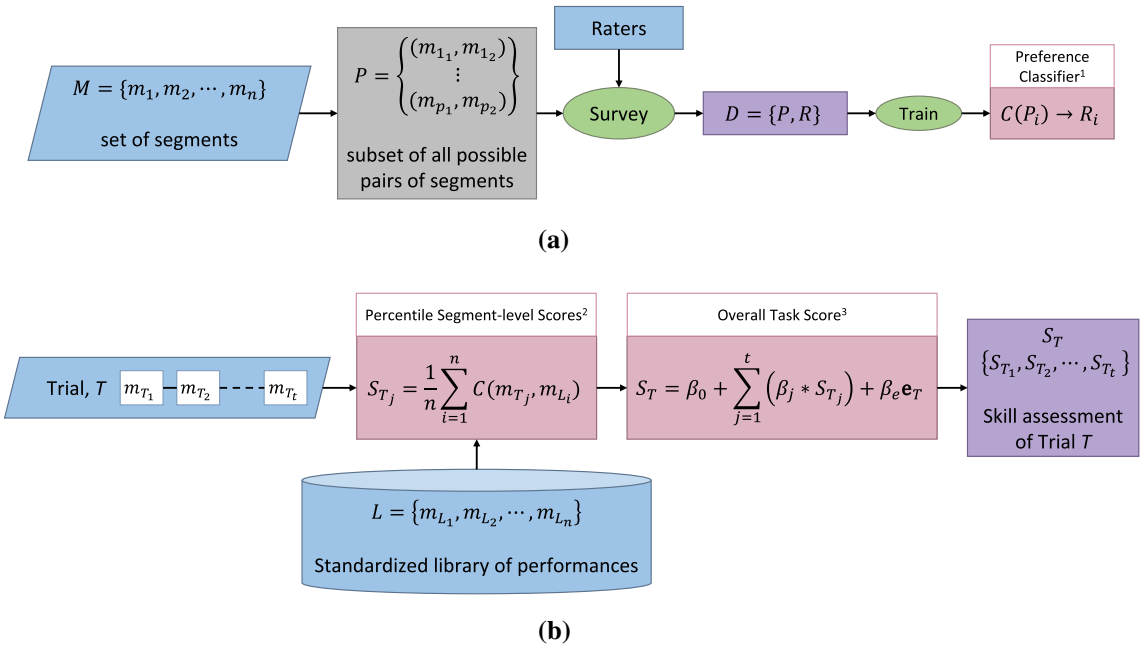

Fig. 1 Components of our framework (shown in *pink blocks*) for objective surgical skill assessment: (1) preference classifier, (2) percentile segment-level scores, and (3) overall task score. **a** The set R represents manual preferences assigned to pairs of segments in the set P by the raters. **b** Given a new instance of a task T, our framework assigns percentile scores to the constituent segments by comparing them against a library of performances L. An overall task-level score S_T is computed using the segment-level scores

pair of segments. We refer to this selection as a *preference*. We denote the preference relation using the symbols $\prec$ and $\succ$ and define it as follows:

$m_1 \prec m_2$ if m_2 is better than m_1
$m_1 \succ m_2$ if m_1 is better than m_2

where m_1 and m_2 are task segment performances. Based on this definition of preference, the binary classifier C is described as below:

$$C(\mathbf{f_1}, \mathbf{f_2}) = \begin{cases} 1 & \text{if } m_1 \succ m_2\,, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{f_i}$ is a feature vector representing the segment-level performance using metrics for surgical skill. We use simple quantitative metrics (listed in Table 1) derived from data on surgical tools and endoscopic camera motion. We compute path length, ribbon area, movements, gripper activations, working distance, console path length, and console workspace separately for the left and right instruments/hands ($2 \times 7 = 14$ features) and two time-based features. Thus, $\mathbf{f_i}$ is a 16 dimensional vector for each task segment. We train the classifier using manually assigned pairwise preferences as the ground-truth labels.

Percentile scores for task segments

The second component of our framework involves computing an objective skill score for individual task segments. Consider a task performance, T consisting of t segments and the jth such segment, m_{T_j}. We apply C to compare m_{T_j} with all instances of a segment performance from a corpus, $\mathscr{L} = \{m_{L_1}, m_{L_2}, \ldots, m_{L_n}\}$ containing n samples (Fig. 1b). Subsequently, we compute the percentile score (S_{T_j}) for m_{T_j} as follows:

$$S_{T_j} = \frac{1}{n}\sum_{i=1}^{n} C(\mathbf{f}_{T_j}, \mathbf{f}_{L_i}) \tag{2}$$

where $\mathbf{f}_{T_j}$ and $\mathbf{f}_{L_i}$ are feature vectors corresponding to the segments m_{T_j} and m_{L_i}, respectively. The percentile score S_{T_j} for instance m_{T_j} is the proportion of pairwise comparisons between m_{T_j} and each instance m_{L_i} in $\mathscr{L}$ where $m_{T_j} \succ m_{L_i}$.

Overall task score

The third component of our framework involves computing an objective measure of surgical skill for the overall task based on automated assessments of the constituent task segments. We hypothesize that a linear summation of the percentile scores for all segments within a task will yield an objective and valid overall task score. Accordingly, we train a

Table 1 Quantitative metrics using instrument and camera motion data from [8–10,16,17]

Metric	Description
Time	Time in seconds to complete the task segment
Time fraction	Fraction of overall task time spent in performing the segment
Path length	Distance traveled by the instrument tip
Ribbon area [8]	Area swept by the instrument shaft
Movements [9,10]	Number of peaks in magnitude of velocity of the instrument tip
Gripper activations	Number of times the instrument gripper was closed
Working distance	Distance between the instrument tip and camera along the view direction
Console path length [16,17]	Distance traveled by the manipulators at the console
Console workspace [16,17]	Bounding volume of the console range of motion

linear regression model to learn parameters for each segment-level score in a task using expert-assigned global rating scores (GRS) as the ground-truth. The model is described below:

$$S_T = \beta_0 + \sum_{j=1}^{t} \beta_j \; S_{T_j} + \beta_e \; \mathbf{e}_T \tag{3}$$

where S_T represents ground-truth GRS for an instance of the task (T), S_{T_j} represents the percentile score for a segment m_{T_j} which was performed as part of the task performance, T (see Fig. 1b). We include $\mathbf{e}_T$ to account for the fraction of total task time spent in portions of the task which did not constitute a semantically meaningful activity segment.

Experiments

For our experiments, we used an existing data set of surgical training task segments. We surveyed a surgically untrained crowd and a group of expert surgeons to obtain the ground-truth[1] for training our preference classifier ("Preference classifier" section).

Surgical task data set

The surgical task data set we used was collected in a previous study [17]. The data set includes instances of a study task (suture throw followed by a surgeon's knot) performed on a bench-top model using the *da Vinci* Surgical System (dVSS, Intuitive Surgical, Inc., Sunnyvale, CA). Four expert and 14 trainee surgeons performed 135 instances of the study task in 45 sessions, with three instances in each session. An expert surgeon watched video recordings for each session and assigned a single GRS using a modified OSATS approach [22]. The expert assessed skill using six criteria, each on a five-point Likert-like scale (with 1 being poor skill and 5 being excellent skill): respect for tissue, time and motion, instrument handling, knowledge of instruments, flow of operation, and knowledge of specific procedure. Thus the overall score has a range of 6 to 30. We applied the session-specific GRS as a task-level skill score for each instance of the study task performed during that session.

The surgical task data set is comprised of: (a) kinematic data describing the motion of the manipulator tips on the patient- and surgeon-sides of the dVSS, (b) stereo endoscopic video recordings, and (c) manual annotations of constituent maneuvers for each instance of the task. Maneuvers represent circumscribed segments or milestones that describe a semantically meaningful portion of a surgical task [21].

Figure 2 shows the flow of maneuvers constituting the study task in our data set. We grouped the maneuvers in our study task into the following five categories to account for variability in how different surgeons performed the study task:

- ST1—suture throw performed in two steps; passing the needle separately through each side of the incision or repair ($n = 60$);
- ST2—suture throw performed in one step; passing the needle through both sides of the incision or repair in a single motion ($n = 104$);
- GPR—running suture out of tissue following a suture throw ($n = 154$);
- KT1—the first knot ($n = 135$);
- KT2—any knot thrown subsequent to the first knot ($n = 203$).

In addition to the maneuver categories listed above, our vocabulary for maneuvers in the study task included inter-maneuver segments (IMS; denoted by green circles in Fig. 2).

[1] The term ground-truth, here and henceforth, has been used to denote a *reference* value obtained by pooling the crowd/expert responses.

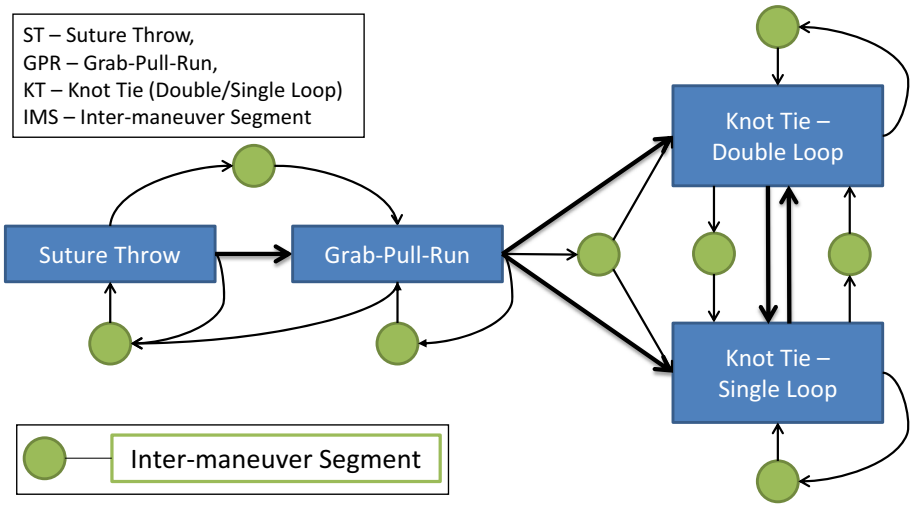

Fig. 2 Maneuver flow in the study task of suturing and knot tying

IMS represent portions of the task wherein the surgeons performed certain actions in preparation for the next maneuver.

Crowdsourcing user study

We conducted a crowdsourcing user study (approved by The Johns Hopkins Homewood Institutional Review Board) to generate two different sources of ground-truth for training the preference classifier in our framework—surgically untrained individuals (crowd), and faculty surgeons (experts). We hosted a survey on a website for the crowd and expert participants to complete the specified human intelligence tasks (HITs), which in our case was to provide preferences for pairs of maneuvers. The study call was voluntary and open to all within the Johns Hopkins community. We generated the HITs by forming pairs of maneuvers belonging to the same category. We did not include IMS when generating HITs because the actions performed across instances of IMS in our data set were highly variable in nature and in the goals they accomplished. The maneuver videos were typically 20–30 s in length.

Based on a priori sample size calculations, we sampled a total of 360 HITs for the crowd and a subset of 120 of those 360 HITs for the experts. We assumed that the proportion of pairs with correct ordering will be 85 % for the crowd and 90 % for the experts. Accordingly, we computed that we will be able to estimate the proportion of pairs with accurate ordering of videos with a 95 % confidence interval (CI) of width of 0.1 (10 %) if we recruited 49 crowd participants and 35 expert participants. Furthermore, we computed the sample size to test a hypothesis of equivalence comparing accuracy of the preference classifiers trained using preferences obtained from crowd and expert participants. We assumed that the accuracy of classifier trained with crowd ratings will be 80 % and accuracy of classifier trained with expert ratings will be 85 %. We estimated that we would have 90 % power to establish equivalence within a 10 % margin with 52 unique pairs of videos.

We grouped HITs into 12 surveys of 30 HITs each for crowd participants, and two surveys of 30 HITs and six surveys of 15 HITs for experts. This division satisfied the required sample size while making the overall length of the surveys shorter to encourage expert participation. A study participant was required to complete all the HITs belonging to a survey in order for their participation to be complete. Additionally, attention HITs, consisting of an obviously good performance versus an obviously poor performance, were presented to the participants at regular intervals (every 10 HITs). Participants who did not provide correct preferences for such HITs were automatically disqualified from the study.

The participants were asked to sign an informed consent for the study on the welcome page and were registered using their name and email address. The participants were allowed to participate in any number of surveys, but only once in each survey. Participants who failed an attention HIT were not allowed to participate in any other survey. Additionally, the crowd participants were provided a compensation of $10 gift card per survey. They were given a period of three days starting from the time they sign up for a particular survey, after which they would be automatically disqualified. The expert participants were given a period of seven days to finish their survey once they signed up for it. There were no restrictions on the amount of time spent by a participant on an individual HIT.

The image in Fig. 3 illustrates a typical screen visualized by study participants. We asked the participants to specify which of the two maneuvers displayed on the screen appeared to have been performed with greater skill (preference), and to specify their level of confidence in choosing the preference (as shown in Fig. 3) on a Likert-like scale. The answer options were enabled, only when the participant had completely viewed both the videos.

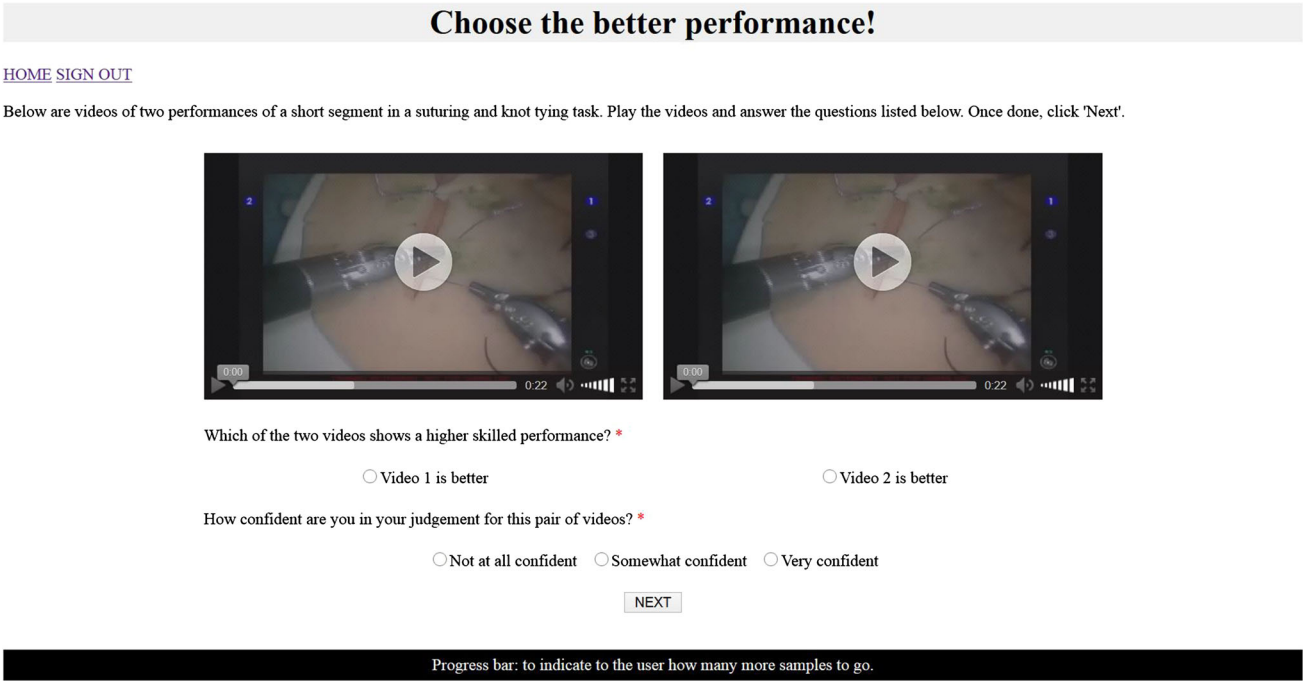

Fig. 3 The Web-based survey page showing a sample HIT

We recruited 147 crowd participants across the 12 surveys, most were students from the engineering, arts and sciences programs at the Johns Hopkins University. We restricted the total number of crowd participants per survey (Survey 1: 52 participants, Surveys 2 through 9: 11 participants, Surveys 10 through 12: 5 participants). We were able to recruit eight expert participants across the eight surveys, all of whom were faculty surgeons at the Johns Hopkins Medical Institutions. We restricted the recruitment to three experts per survey to get multiple responses for each of the 120 HITs sampled. We obtained preferences from all the crowd participants within a period of three days, whereas it took about four weeks to capture preferences from the experts. For this reason, we were not able to recruit the number of experts suggested by our power analysis, although, as we note later, the consistency of the experts suggests our analysis was overly conservative. The time spent (in seconds) per HIT[2] across the 120 overlapping HITs were: experts (mean 117.36, σ 230.52), and crowd (mean 71.52, σ 87.91).

HIT agreement and HIT confidence

For each HIT, we computed two properties *viz.* agreement and confidence. We computed the agreement (agr) property as the percentage of participants completing the HIT that gave the same preference with a confidence level of five as defined in Eq. 4 below:

$$\text{agr}_h = \max\left(\frac{r_h}{k_h}, \frac{k_h - r_h}{k_h}\right) \tag{4}$$

where k_h is the total number of participants who provided their preference rating for the HIT h, with a confidence level of 5, r_h is the number of participants preferring one segment among the pair presented in the HIT. To ensure that our preference classifier was trained on a meaningful ground-truth, we used only those HITs for training where agr ≥ 0.75.

Another characteristic property of the HITs is the confidence (conf), which was computed as an average of confidence level weights (Table 2) assigned by participants responding to that HIT as shown in Eq. 5.

$$\text{conf}_h = \frac{1}{k_h}\sum_{j=1}^{k_h} w_{hj} \tag{5}$$

where w_{hj} is the confidence weight (Table 2) associated with the confidence level indicated by the participant j for their preference for the HIT h, and k_h is the total number of participants who performed the HIT h. By doing so, the classifier was trained using data where the raters were more confident about their preferences. We used HITs with conf ≥ 0.5 in our sensitivity analysis.

[2] The participants could take breaks and come back and answer these HITs at a later time. Thus, we cannot draw any reliable conclusions based on these numbers.

Table 2 Confidence levels elicited in the survey and corresponding weights for ratings

Survey phrase	Level	Weight
Very confident	5	1.0
Somewhat confident	3	0.5
Not at all confident	1	0.0

Pooled preferences from the participants

To obtain a single ground-truth preference per HIT (pair of segments), we investigated three different approaches for majority pooling and one approach for weighted pooling.

In the first approach, we simply selected the majority rating (R_{all}) from all the preference ratings obtained for a given HIT. In the second approach, we selected the majority among ratings where the confidence level was at least three (R_3). In the third approach, we selected the majority among ratings where the confidence level was five (R_5). We used all three approaches for reliability analyses, but only the simple majority rating approach (R_{all}) for validity analyses due to sample size limitations with the remaining majority pooling approaches.

In the weighted pooling approach, we selected the preference using a weighted count of preference ratings for a given HIT (R_{w}). Table 2 shows the weights we used for each level of confidence associated with the ratings. Ratings with confidence level 5 contributed a full count and those with confidence level 3 contributed one-half of a count toward the preference ratings. Ratings with confidence level 1 did not contribute to the preference rating.

Reliability and validity of manually annotated preferences

We evaluated the inter-participant reliability of preferences separately for the crowd and experts using the Fleiss' kappa (κ), which is a standard measure of agreement when multiple participants provide ratings on multiple tasks [12]. Fleiss' kappa represents the agreement beyond what is expected due to chance. A value of $\kappa = 1$ indicates perfect agreement and $\kappa \leq 0$ indicates no agreement or disagreement among raters. We also evaluated validity of preferences obtained from the crowd assuming preferences obtained from the experts were the ground-truth. We computed the percentage agreement or accuracy as the measure of validity. Additionally, we computed the Fleiss' kappa statistic for the confidence level ratings assigned by the crowd and expert participants. We compared the agreement within crowd and expert groups in selecting the majority confidence rating. For this, a metric similar to HIT agreement property (agr) defined in "HIT agreement and HIT confidence" section was calculated as in Eq. 6:

$$\text{agr}_h = \max\left(\frac{r_1}{k_h}, \frac{r_3}{k_h}, \frac{r_5}{k_h}\right) \tag{6}$$

where k_h is the total number of participants who provided their preference rating for the HIT h; r_i is the number of participants who selected their confidence level for the rating to be i for the HIT h.

Validity of preference classifiers

We trained two separate linear support vector machines (SVM; [7]), one using preferences from the crowd and the other from experts. We explored an AdaBoost classifier using stump-based weak learners as well. However, the SVMs performed better than the boosted classifier and thus further analyses were performed using SVMs. We trained each of these SVMs using two different sets of features; the first set (SVM7) matched the 7-D feature vector used in [21] for comparison [time, path lengths (2x), ribbon areas (2x), and movements (2x)], and the second set (SVM16) included the 16 dimensions described in Table 1. We trained a separate classifier for each category of maneuvers ("Surgical task data set"), as well as one overall classifier for all categories of maneuvers pooled together. In addition, we trained separate classifiers for preferences obtained with two pooling approaches - R_{all} and R_{w} ("Pooled preferences from the participants"). We evaluated crowd- and expert-based preference classifiers against the respective manually assigned preferences as the ground-truth. We used a tenfold cross-validation approach and computed accuracy between the classifier-assigned preferences and participant-assigned preferences.

We computed the accuracy of the crowd preference classifier while varying the number of training samples used. A fraction (20 %) of the HITs was held out as a fixed test data set. The number of training samples (n) was incremented in steps of 10 samples at a time. For each n, an average accuracy was calculated using 20 bootstrap iterations for sampling the training data.

Validity of our framework for objective skill assessment

We compared the task-level scores obtained using the expert preference classifier against ground-truth GRS. We trained a simple linear regression model (Eq. 3) to predict the ground-truth GRS in a leave-one-out cross-validation approach. The predictors for the model included the segment-level scores as a four-dimensional vector (ST, GPR, KT1, KT2), the number of IMS, fraction of total task time spent performing IMS, and the fraction of total task time that was not annotated with any maneuver label. The latter three terms in the predictors formed $\mathbf{e}_T$ from Eq. 3. The segment score for ST was obtained

from the score for ST1 or ST2, whichever was performed in the given instance of the task.

We computed the root-mean-squared error (RMSE) and the Spearman's correlation coefficient (ρ) between predicted and ground-truth scores as measures of validity. The Spearman's correlation is a nonparametric measure of association between two ranked variables. A value of $\rho = +1$ indicates perfect monotonic dependence, while a value of zero indicates no correlation. In addition, we learned similar regressions to predict scores for each of the six individual components within GRS [22] listed in the "Surgical task data set" section.

Comparison of crowd and expert preference classifiers

We assessed the crowd and expert preference classifiers for three outputs of our framework pipeline:

Accuracy We tested the equivalence of the crowd and expert preference classifiers by checking whether the accuracy of the crowd preference classifier is within the 10 % margin of accuracy of the expert preference classifier. For hypothesis testing purposes, we performed cross-validation using the set of HITs rated by both the crowd and the experts ($n = 75$),[3] while training the respective classifiers using all of the held out data available per group of users. Additionally, we performed a sensitivity analysis using only those HITs rated by both the crowds and experts for training as well as testing in a leave-one-out cross-validation approach. More training data were available for the crowd classifier as compared to the expert classifier in the former analysis, whereas the training data for the two classifiers remained fixed in the latter case.

Segment-level scores We computed a Spearman's correlation coefficient between the segment-level scores obtained from the crowd and expert preference classifiers, separately for each maneuver category.

Task-level scores We computed a Pearson's correlation coefficient (ρ) between the task-level scores obtained using the crowd and expert preference classifiers. The Pearson's correlation measures the linear correlation between two continuous variables. A value of +1 for the Pearson's correlation indicates total positive correlation, 0 indicates no correlation, and −1 indicates total negative correlation. In addition, we tested whether the task-level scores obtained using the crowd and expert preference classifiers were statistically equivalent to each other within a prespecified margin of two units on the GRS scale.

[3] The number of HITs rate by both the crowd and experts was 120. However, filtering the HITs based on the agreement metric ("HIT agreement and HIT confidence") drops the count to 75.

Results

Reliability and validity of manually annotated preferences

As shown in Table 3, we observed moderate inter-rater agreement within both the crowd and expert participants. Experts appeared to have a higher inter-rater agreement compared with the crowd, as one would expect.

The crowd preferences were at least 83 % accurate when taking expert preferences as the ground-truth. This accuracy was robust across all four approaches for pooling preferences ("Pooled preferences from the participants") for a given HIT, as shown in Table 4. The accuracy increased with the R_3 and R_5 pooling approaches, as one would expect with ratings having higher confidence.

Inter-participant agreement seemed to be higher for ratings with higher confidence levels for both the crowd and experts, as shown in the Table 5. However, the agreement (based on Fleiss' kappa) within the group of participants on their confidence level rating was observed to be very low −0.08 (crowd) and 0.22 (experts).

Table 3 Inter-participant reliability for crowdsourced preferences using percentage agreement (agr) and Fleiss' kappa (κ)

Group	# HITs	# Workers	agr	95 % CI (agr)	κ	95 % CI (κ)
Crowd	360	147	0.81	(0.80, 0.83)	0.41*	(0.40, 0.42)
Expert	120	8	0.88	(0.85, 0.91)	0.55*	(0.45, 0.64)

* P value < 0.001

Table 4 Agreement between pooled preferences for HITs which were rated by both the crowd and expert participants

Statistic	Pooling approach			
	R_{all}	R_3	R_5	R_w
# HITs	120	118[a]	89[a]	120
perc agr	0.83	0.85	0.87	0.84
95 % CI	(0.77, 0.90)	(0.78, 0.91)	(0.79, 0.94)	(0.78, 0.91)

perc agr percentage agreement

[a] HITs that did not get ratings with confidence levels of at least 3 and at least 5, respectively, were omitted while computing the above statistical measures

Table 5 Inter-participant agreement for crowdsourced confidence levels using agreement (agr) property (Eq. 6)

Confidence level	1	3	5
Crowd	0.499	0.577	0.652
Expert	0.535	0.728	0.848

Validity of preference classifiers

A preference classifier trained using ratings obtained from the crowd was able to predict the crowd's pooled preferences with an accuracy of 85 % (SE 2 %). The preference classifier trained by expert preferences had an accuracy of 89 % (SE 3 %). As noted before in Table 3, the crowd participants agreement across the HITs was 81 with a 95 % confidence interval of (80,83), while the experts had an agreement of 88 % with a 95 % CI of (85,91). Thus, the performance of our classifier is above par compared to the inter-observer agreement.

The accuracy of the crowd preference classifier improved when the training data were filtered to only include HITs with an overall confidence of 0.5 or more (see Table 6). But this was not the case with the expert preference classifier, where the accuracy appeared to decrease when we filtered the training data to include HITs with an overall confidence of 0.5 or more. Extending the set of training features did not appear to consistently improve accuracy of either the crowd or expert preference classifier.

Accuracy of the preference classifiers did not appear to be sensitive to whether we pooled preferences using R_{all} or R_{w}. Accuracy for the expert preference classifier for R_{all} was consistently greater than those for R_{w}, but the difference was small in magnitude. We did not observe a consistent direction for these differences with the crowd preference classifier (see Table 6).

Table 6 also shows that classifiers specific to some maneuver categories (KT1 and KT2) appeared to be more accurate than the overall classifier in predicting manual preferences. This was not the case for classifiers specific to other maneuver categories (ST1, ST2, and GPR).

The average accuracy of the crowd preference classifier trained using a varying number of training samples is shown in Fig. 4. The accuracy plateaus after $n = 120$ training samples with a value of 0.80 showing a change in the order of 0.2 as the number of training samples varies in the range of (120, 220). We did not conduct a similar analysis for the expert preference classifier due to a small sample size.

Validity of our framework for objective skill assessment

Using the expert preference classifier, we predicted expert-assigned overall GRS with RMSE lower than one standard deviation (σ) of the ground-truth (RMSE = 5.54; 0.85 σ). The Spearman's correlation between the predicted and ground-truth GRS was 0.55 (P value <0.001).

For components within GRS, the RMSE was 1.05 for respect for tissue, 0.95 for time and motion, 1.16 for instrument handling, 1.01 for knowledge of instruments, 1.20 for flow of operation, and 1.14 for knowledge of specific proce-

Table 6 Accuracies for preference classifiers with crowd and expert preferences

Pooling	Segment	HITs (agr ≥ 0.75)			HITs (agr ≥ 0.75, conf ≥ 0.5)		
		N	SVM7	SVM16	*N*	SVM7	SVM16
Preference classifier trained using crowd preferences							
R_{all}	ST1	30	0.73 (0.08)	0.40 (0.09)	20	0.75 (0.10)	0.65 (0.11)
	ST2	53	0.74 (0.06)	0.70 (0.06)	46	0.76 (0.06)	0.78 (0.06)
	GPR	54	0.78 (0.06)	0.74 (0.06)	41	0.78 (0.06)	0.78 (0.06)
	KT1	62	0.92 (0.03)	0.85 (0.04)	60	0.92 (0.04)	0.85 (0.05)
	KT2	78	0.88 (0.04)	0.88 (0.04)	73	0.90 (0.03)	0.89 (0.04)
R_{all}	ALL	277	0.81 (0.02)	0.82 (0.02)	240	0.82 (0.02)	0.85 (0.02)
R_{w}	ALL	277	0.81 (0.02)	0.80 (0.02)	240	0.85 (0.02)	0.86 (0.02)
Preference classifier trained using expert preferences							
R_{all}	ST1[b]	–	–	–	–	–	–
	ST2	15	0.47 (0.13)	0.80 (0.10)	14	0.50 (0.13)	0.71 (0.12)
	GPR	20	0.85 (0.08)	0.75 (0.10)	20	0.90 (0.07)	0.75 (0.10)
	KT1	25	0.88 (0.06)	0.92 (0.05)	23	0.87 (0.07)	0.87 (0.07)
	KT2	26	0.92 (0.05)	1.00 (0.00)	24	0.83 (0.08)	0.96 (0.04)
R_{all}	ALL	89	0.89 (0.03)	0.89 (0.03)	84	0.85 (0.04)	0.86 (0.04)
R_{w}	ALL	89	0.87 (0.04)	0.87 (0.04)	84	0.83 (0.04)	0.85 (0.04)

Training data were filtered using agreement (agr) and confidence (conf) of a HIT as defined in the "HIT agreement and HIT confidence" section. *N* is the number of HITs available for cross-validation after the filtering. The numbers are reported as accuracies in predicting the manual preference (standard errors)

[a] SVM7 was trained using a subset of metrics listed in Table 1 to match our previous work [21], SVM16 was trained using all the metrics

[b] *N* was too low to perform cross-validation

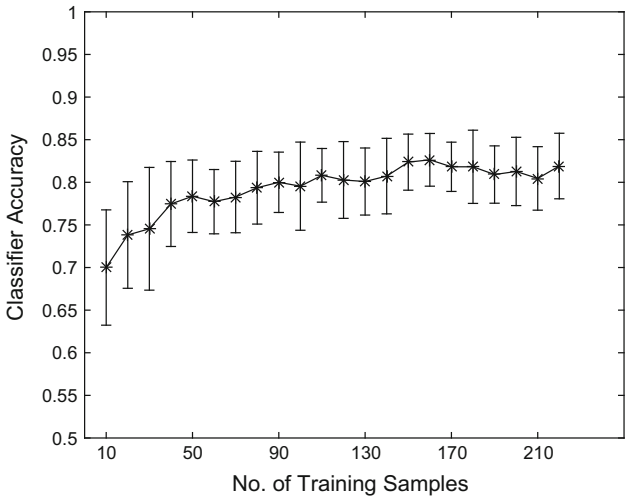

Fig. 4 Crowd preference classifier accuracy versus the number of training samples available. The points on the plot are mean accuracy over a bootstrap sampling of 20 iterations for each setting of the number of training samples. The *error bars* indicate the standard deviation in the accuracy of the classifier

dure. The corresponding Spearman's correlations were 0.52, 0.56, 0.53, 0.63, 0.45, and 0.33, respectively. The correlation coefficients for all the components were statistically significant.

Comparison of crowd and expert preference classifiers

Accuracy As shown in Fig. 5a, our analyses did not demonstrate equivalence between the crowd and expert preference classifiers within a margin of 10 %. Our observation is consistent for SVM7 and SVM16 using training data obtained with different pooling approaches and filtered based on confidence property of the HITs. Using the same training data for the crowd and expert preference classifiers did not alter the outcome of the analysis, as can be seen in Fig. 5b.

Segment-level scores In the case of SVM7, segment-level scores obtained using the crowd preference classifier were highly correlated with those from the expert preference classifier ($\rho \geq 0.86$ for all maneuver categories). But in the case of SVM16, the correlation between the segment-level scores from the two preference classifiers was very sensitive to the sample size specific to the maneuver category. The correlation coefficient was as low as 0.11 for ST1 and as high as 0.85 for KT1.

Task-level scores Task-level scores predicted using segment-level scores from the crowd preference classifier were also highly correlated with those from the expert preference classifier ($\rho \geq 0.84$). As shown in Fig. 5c, the task-level scores obtained using the crowd preference classifier were statistically equivalent to those obtained using the expert preference classifier within a margin of two units on the GRS scale.

Discussion

Our findings in this study are strongly supportive of our framework for objective surgical skill assessment using pairwise comparisons of task segments. Our data indicate that assessments of segment-level skill can be obtained with moderate reliability from surgically untrained individuals as well as from expert surgeons. Further, we show that crowdsourcing is an efficient, reliable, and valid solution for assessing surgical skills at the segment-level. The crowd yielded preferences for maneuvers with high validity when compared with expert surgeons (Table 4), and within three days compared with about four weeks for experts. The experts in our sample were affiliated with various surgical divisions and represented a wide range of experience (number of years in practice). Given the agreement among these diverse experts that we observed in our sample, we expect that our findings will be robust to ground-truth specified by a larger group of experts.

Accuracy of manual preferences by the crowd translated directly into validity of all aspects of our framework. Given ground-truth pairwise preferences for task segments, we demonstrated that a classifier can be trained with sufficient accuracy to yield valid and objective skill assessments at both the segment- and task-levels (Table 6). We did not observe a consistent improvement in the accuracy of the preference classifier by extending the set of features from SVM7 to SVM16. Even though the accuracy for the crowd and expert preference classifiers was not equivalent, both segment- and task-level scores obtained from the two classifiers were highly comparable (Fig. 5). Furthermore, our framework yielded task-level GRS with an error that is comparable in magnitude to the variability we observed in our data set for task-level GRS assigned by an expert surgeon.

Our study establishes a basis for evaluating the educational value of targeted feedback based upon segment-level skill assessment. Segment-level assessments obtained from our framework can be used to provide trainees with targeted feedback on where in the task they need to perform better. Such targeted feedback may facilitate deliberate practice and consequently, effective and efficient skills acquisition. Targeted feedback in the form of coaching by a mentor has been shown to reduce errors in performance and improve skill acquisition [6]. Our framework may also be usefully deployed for standardized evaluation of acquisition, maintenance, and retention of technical skills in the training laboratory. Using a common library of maneuver performances that span a wide spectrum of surgical skill (novice to expert) allows standardized evaluation of trainees across institutions and over time.

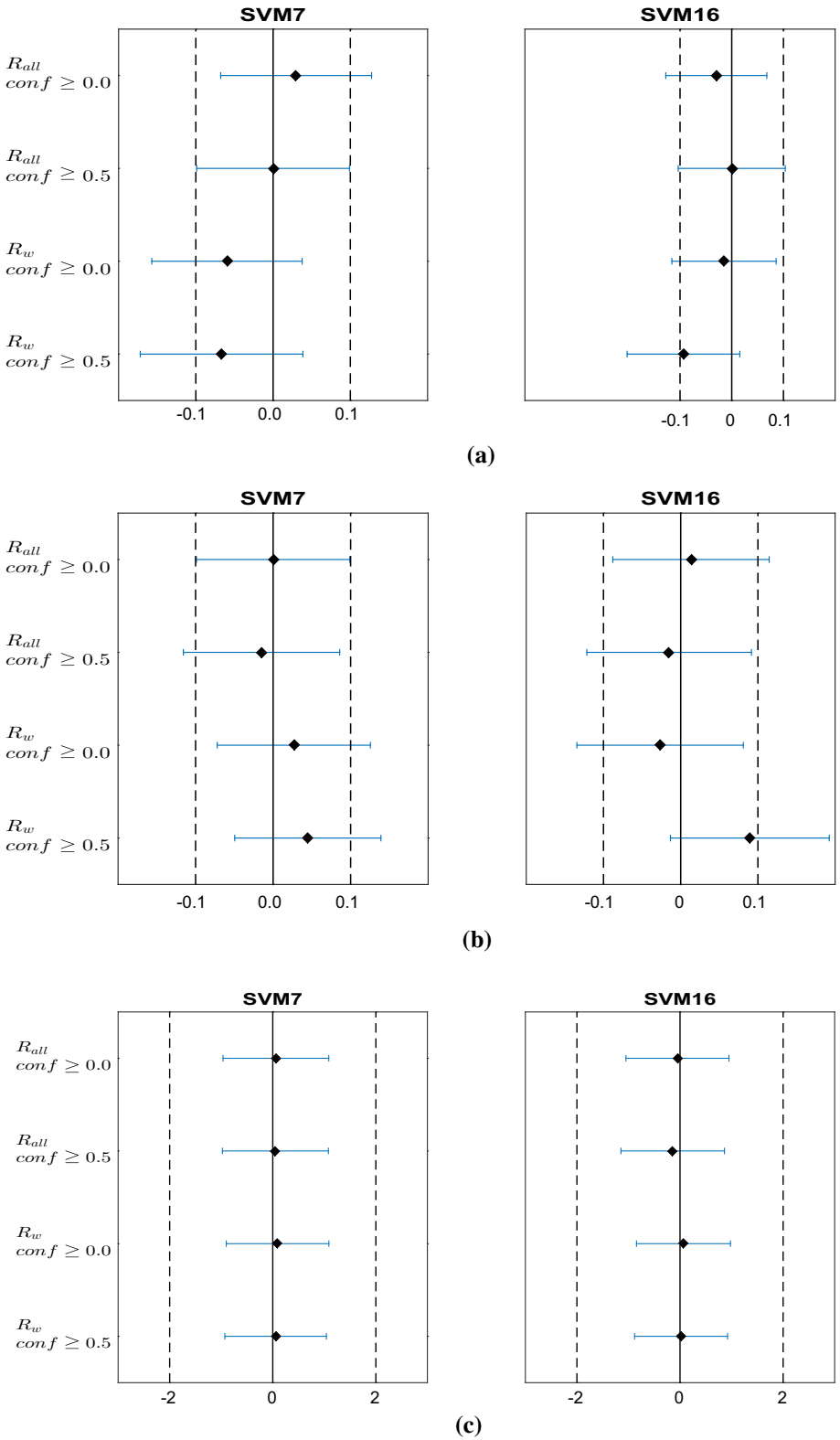

Fig. 5 Equivalence testing of the crowd and expert preference classifiers. The X-axis is the difference in property/outcome measure from the crowd and expert preference classifiers. The *dashed lines* illustrate the equivalence margin on either side of the null value (*solid line*). The *solid diamonds* represent the estimate of the difference in property/outcome obtained from the two classifiers. The *horizontal bars* are the 95 % confidence intervals (CI) for the estimates. Equivalence holds if the 95 % CI lie entirely within the region bounded by the *dashed lines*. **a** Accuracy of preference classifiers using all available training data. **b** Accuracy of preference classifiers using common training data. **c** Task-level scores obtained from the preference classifiers

Surgeons and educators acknowledge the need for such standardization of training and evaluation [3]. Finally, we note that our approach may be deployed on any surgical platform where we can capture the data necessary to compute quantitative measures of surgical skill. This includes robotic, open, conventional laparoscopic, and endoscopic surgery. We used only tool motion data to compute features to train the preference classifiers, but other sources of data such as video

images may also be used for this purpose either alone or in combination with each other. For example, Ahmidi et al. capture motion data in an open procedure to preform reliable skill assessment in [2].

One remaining limitation of this work is the fact that our approach requires prior segmentation of the study tasks into constituent segments. This assumes both that such constituent segments exist and that the resources or infrastructure to perform this segmentation exist. While crowdsourcing annotation of segments within a task is, in principle, possible, the reliability of such an approach has yet to be established. Several tools have been developed for automatic segmentation of tasks into finer segments (gestures), but none exist for segmentation of tasks into maneuvers [1,15,27,29,31]. Finally, we studied a single surgical task, suturing and knot tying, performed on the robotic surgical platform. Further studies validating our framework may focus on other tasks within typical surgical skills training curricula performed using non-robotic surgical platforms.

An interesting and open question is whether pairwise comparisons provide a more effective means for crowdsourced skill assessment than global assessments, and whether the effectiveness of the framework is sensitive to the granularity of analysis. Conversely, the most effective level of analysis for teaching is also not yet established. Feedback at levels finer than maneuvers in the task, such as gestures, may be important for surgical skills acquisition. For example, errors in performance of the task are typically articulated at the gesture-level, and thus, gesture-level assessments using our framework may yield effective feedback for trainees. The effectiveness or educational value of gesture-, maneuver-, and task-level assessment for acquisition, maintenance, and retention of surgical technical skills remains to be investigated in future studies. We also note that technical skills is one component of the overall performance in the operating room, and further work to incorporate preoperative and postoperative skills can help predict patient outcomes.

Conclusion

We have presented a framework for crowdsourced skill assessment that yields valid objective surgical skill assessments both for the overall task and for maneuvers within a task. We have shown that crowdsourcing can provide reliable pairwise comparisons for maneuvers within a task and that pairwise comparisons by a surgically untrained crowd used within our framework yield segment- and task-level assessments that are comparable to those obtained using pairwise comparisons by expert surgeons.

Acknowledgments We acknowledge all participants in our crowdsourcing user study, and Intuitive surgical, Inc., for facilitating capture of data from the dVSS. A combined effort from the Language of Surgery project team led to the development of the manual task segmentation. The Johns Hopkins Science of Learning Institute and internal funding from the Johns Hopkins University supported this work.

Compliance with ethical standards

Conflict of interest Anand Malpani, S Swaroop Vedula, C C Grace Chen, and Gregory D Hager declare that they have no conflict of interest.

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

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

References

1. Ahmidi N, Gao Y, Bjar B, Vedula SS, Khudanpur S, Vidal R, Hager GD (2013) String Motif-Based Description of Tool Motion for Detecting Skill and Gestures in Robotic Surgery. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N (eds.) Medical image computing and computer-assisted intervention MICCAI 2013, no. 8149 in Lecture Notes in Computer Science. Springer, Berlin Heidelberg, pp 26–33. http://link.springer.com/chapter/10.1007/978-3-642-40811-3_4
2. Ahmidi N, Poddar P, Jones JD, Vedula SS, Ishii L, Hager GD, Ishii M (2015) Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. Int J Comput Assist Radiol Surg. doi:10.1007/s11548-015-1194-1. http://link.springer.com/article/10.1007/s11548-015-1194-1
3. Bell Jr RH (2009) Why Johnny cannot operate. Surg 146(4):533–542. doi:10.1016/j.surg.2009.06.044. http://www.sciencedirect.com/science/article/pii/S0039606009004620
4. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442. doi:10.1056/NEJMsa1300625. http://www.nejm.org/doi/full/10.1056/NEJMsa1300625
5. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, Lendvay T (2014) Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 187(1):65–71. doi:10.1016/j.jss.2013.09.024. http://www.sciencedirect.com/science/article/pii/S0022480413008998
6. Cole SJ, Mackenzie H, Ha J, Hanna GB, Miskovic D (2014) Randomized controlled trial on the effect of coaching in simulated laparoscopic training. Surg Endosc 28(3):979–986. doi:10.1007/s00464-013-3265-0. http://link.springer.com/article/10.1007/s00464-013-3265-0
7. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. doi:10.1007/BF00994018. http://link.springer.com/article/10.1007/BF00994018
8. Curet M, Dimaio SP, Gao Y, Hager GD, Itkowitz B, Jog AS, Kumar R, Liu M (2012) Method and system for analyzing a task trajectory. International Classification A61B19/00, G01C21/00; Cooperative Classification A61B19/2203, G01C21/00, A61B19/00
9. Datta V, Chang A, Mackay S, Darzi A (2002) The relationship between motion analysis and surgical technical assessments. Am J Surg 184(1):70–73. doi:10.1016/S0002-9610(02)00891-7. http://www.sciencedirect.com/science/article/pii/S0002961002008917

10. Dosis A, Aggarwal A, Bello F, Moorthy K, Munz Y, Gillies D, Darzi A (2005) Synchronized video and motion analysis for the assessment of procedures in the operating theater. Arch Surg 140(3):293–299. doi:10.1001/archsurg.140.3.293. http://dx.doi.org/10.1001/archsurg.140.3.293
11. Dwork C, Kumar R, Naor M, Sivakumar D (2001) Rank aggregation methods for the web. WWW '01. ACM, New York, NY, USA. pp 613–622. doi:10.1145/371920.372165. http://doi.acm.org/10.1145/371920.372165
12. Fleiss JL, Levin B, Paik MC (2003) The measurement of interrater agreement. In: Statistical methods for rates and proportions. Wiley, pp 598–626. http://onlinelibrary.wiley.com/doi/10.1002/0471445428.ch18/summary
13. Freund Y, Iyer R, Schapire R, Singer Y (2003) An efficient boosting algorithm for combining preferences. J Mach Learn Res 4:933–969. http://dl.acm.org/citation.cfm?id=945365.964285
14. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187(1):247–252. doi:10.1016/j.juro.2011.09.032
15. Haro BB, Zappella L, Vidal R (2012) Surgical gesture classification from video data. In: Ayache N, Delingette H, Golland P, Mori K (eds.) Medical image computing and computer-assisted intervention MICCAI 2012. Springer, Berlin, pp 34–41. http://link.springer.com/chapter/10.1007/978-3-642-33415-3_5
16. Kumar R, Jog A, Malpani A, Vagvolgyi B, Yuh D, Nguyen H, Hager G, Chen C (2012) Assessing system operation skills in robotic surgery trainees. Int J Med Rob Comput Assist Surg 8(1):118–124. doi:10.1002/rcs.449. http://onlinelibrary.wiley.com/doi/10.1002/rcs.449/abstract
17. Kumar R, Jog A, Vagvolgyi B, Nguyen H, Hager G, Chen CCG, Yuh D (2012) Objective measures for longitudinal assessment of robotic surgery training. J Thorac Cardiovasc Surg 143(3):528–534. doi:10.1016/j.jtcvs.2011.11.002. http://www.sciencedirect.com/science/article/pii/S0022522311012748
18. Kumar R, Rajan P, Bejakovic S, Seshamani S, Mullin G, Dassopoulos T, Hager G (2009) Learning disease severity for capsule endoscopy images. pp 1314–1317. doi:10.1109/ISBI.2009.5193306
19. Maier-Hein L, Mersmann S, Kondermann D, Bodenstedt S, Sanchez A, Stock C, Kenngott HG, Eisenmann M, Speidel S (2014) Can masses of non-experts train highly accurate image classifiers? In: Golland P, Hata N, Barillot C, Hornegger J, Howe R (eds.) Medical Image Computing and Computer-Assisted Intervention MICCAI 2014, no. 8674 in Lecture Notes in Computer Science. Springer International Publishing, pp 438–445. http://link.springer.com/chapter/10.1007/978-3-319-10470-6_55
20. Maier-Hein L, Mersmann S, Kondermann D, Stock C, Kenngott HG, Sanchez A, Wagner M, Preukschas A, Wekerle AL, Helfert S, Bodenstedt S, Speidel S (2014) Crowdsourcing for reference correspondence generation in endoscopic images. In: Golland P, Hata N, Barillot C, Hornegger J, Howe R (eds.) Medical image computing and computer-assisted intervention MICCAI 2014, no. 8674 in Lecture Notes in Computer Science. Springer International Publishing, pp 349–356. http://link.springer.com/chapter/10.1007/978-3-319-10470-6_44
21. Malpani A, Vedula SS, Chen CCG, Hager GD (2014) Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In: Stoyanov D, Collins DL, Sakuma I, Abolmaesumi P, Jannin P (eds.) Information processing in computer-assisted interventions. Springer International Publishing, pp 138–147. http://link.springer.com/chapter/10.1007/978-3-319-07521-1_15
22. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84(2):273–278
23. Reiley CE, Hager GD (2009) Task versus subtask surgical skill svaluation of robotic minimally invasive surgery. In: Yang GZ, Hawkes D, Rueckert D, Noble A, Taylor C (eds.) Medical image computing and computer-assisted intervention MICCAI 2009. Springer, Berlin, pp 435–442. http://link.springer.com/chapter/10.1007/978-3-642-04268-3_54
24. Rosen J, Hannaford B, Richards C, Sinanan M (2001) Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. IEEE Trans Biomed Eng 48(5):579–591. doi:10.1109/10.918597
25. Rosen J, Solazzo M, Hannaford B, Sinanan M (2002) Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden Markov model. Comput Aided Surg 7(1):49–61. doi:10.1002/igs.10026. http://onlinelibrary.wiley.com/doi/10.1002/igs.10026/abstract
26. Sharma Y, Plotz T, Hammerld N, Mellor S, McNaney R, Olivier P, Deshmukh S, McCaskie A, Essa I (2014) Automated surgical OSATS prediction from videos, pp 461–464. doi:10.1109/ISBI.2014.6867908
27. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R (2012) Sparse hidden Markov models for surgical gesture classification and skill evaluation. In: Abolmaesumi P, Joskowicz L, Navab N, Jannin P (eds.) Information processing in computer-assisted interventions. Springer, Berlin, pp 167–177. http://link.springer.com/chapter/10.1007/978-3-642-30618-1_17
28. Van Eaton EG, Tarpley JL, Solorzano CC, Cho CS, Weber SM, Termuhlen PM (2011) Resident education in 2011: Three key challenges on the road ahead. Surgery 149(4):465–473. doi:10.1016/j.surg.2010.11.007. http://www.sciencedirect.com/science/article/pii/S0039606010006148
29. Varadarajan B, Reiley C, Lin H, Khudanpur S, Hager G (2009) Data-derived models for segmentation with application to surgical assessment and training. In: Yang GZ, Hawkes D, Rueckert D, Noble A, Taylor C (eds.) Medical image computing and computer-assisted intervention. Springer, Berlin, pp 426–434. http://link.springer.com/chapter/10.1007/978-3-642-04268-3_53
30. Vassiliou M, Feldman L, Andrew C, Bergman S, Leffondr K, Stanbridge D, Fried G (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190(1):107–113. doi:10.1016/j.amjsurg.2005.04.004
31. Zappella L, Bjar B, Hager G, Vidal R (2013) Surgical gesture classification from video and kinematic data. Med Image Anal 17(7):732–745. doi:10.1016/j.media.2013.04.007. http://www.sciencedirect.com/science/article/pii/S1361841513000522