Int J CARS (2017) 12:1151–1159
DOI 10.1007/s11548-017-1610-9

ORIGINAL ARTICLE

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

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

Received: 27 January 2017 / Accepted: 8 May 2017 / Published online: 18 May 2017

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

✉ Rodney L. Dockter
dockt036@umn.edu

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

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

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

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

Introduction

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

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

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

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

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

Methods

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

Data set

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

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

Table 1 FLS trials by task and skill level

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

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

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

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

Value of 'raw motion data' for classification

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

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

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

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

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

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

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

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

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

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

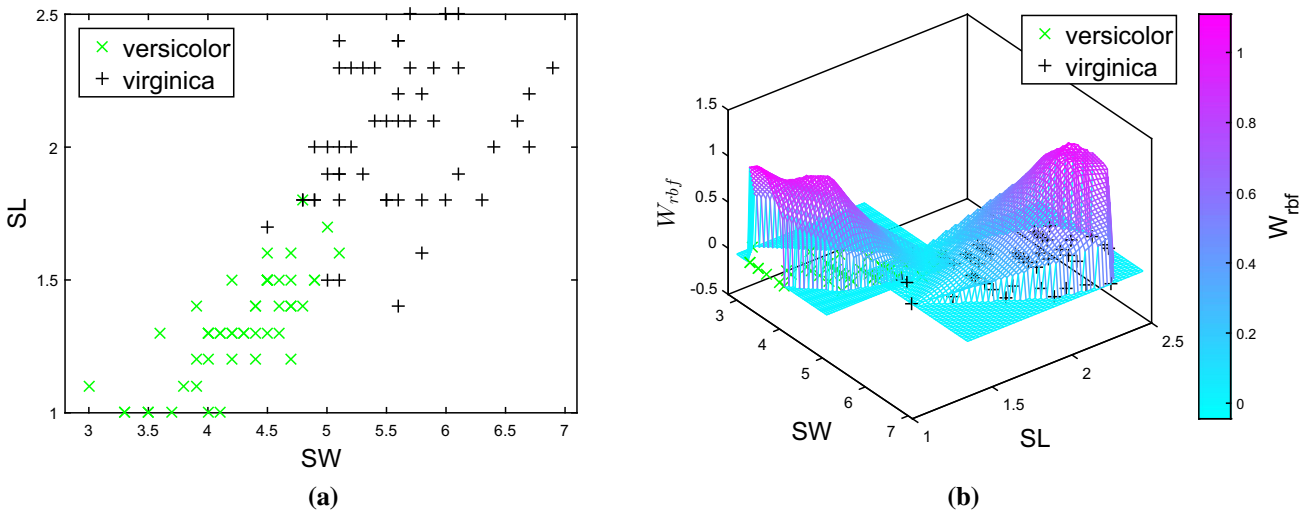

Fig. 1 RELIEF-RBF for sepal width (SW) and sepal length (SL) data from the Versicolor and Virginica classes. **a** 2D Fisher Iris data. **b** RELIEF-RBF weights

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

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

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

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

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

Intent vectors

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

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

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

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

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

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

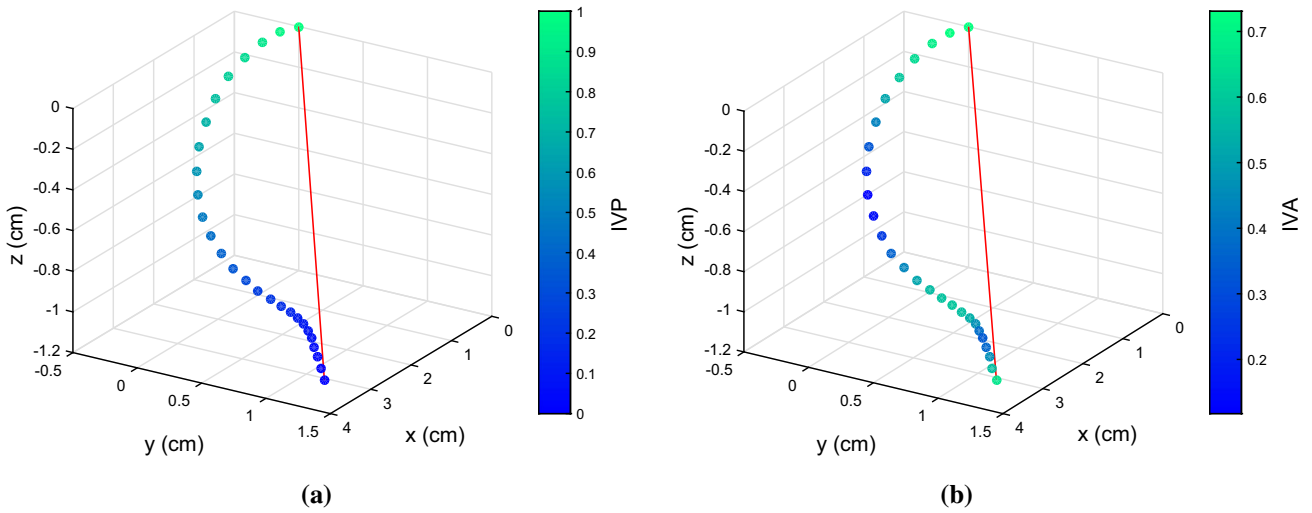

Fig. 2 Intent vector measures. **a** Intent vector progress in 3D. **b** Intent vector angle in 3D

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

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

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

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

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

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

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

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

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

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

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

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

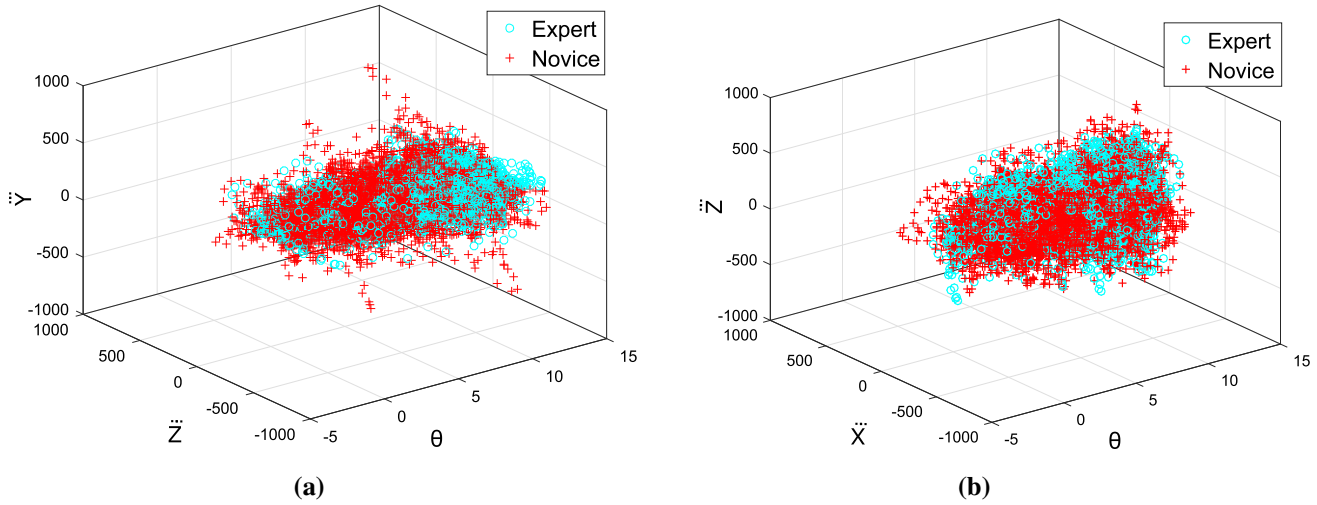

Fig. 3 Relevance weightings for raw motion states. **a** Top three RELIEFF states. **b** Top three RELIEF-RBF states

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

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

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

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

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

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

Results

Value of 'raw motion data' for classification

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

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

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

Intent vectors

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

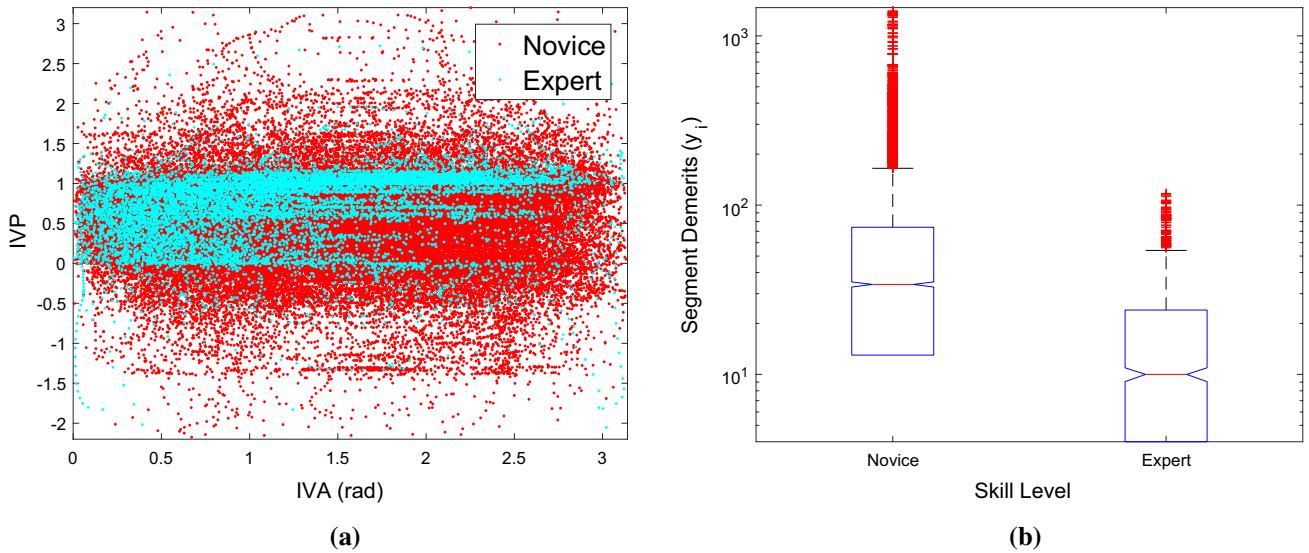

Fig. 4 Intent vector data (**a**) and demerit counts (**b**) (obvious novice and expert) for suturing task box-plot notch indicate range of 95% confidence for median separation. **a** IVA versus IVP with class labels. **b** Per-segment demerits (y_i)

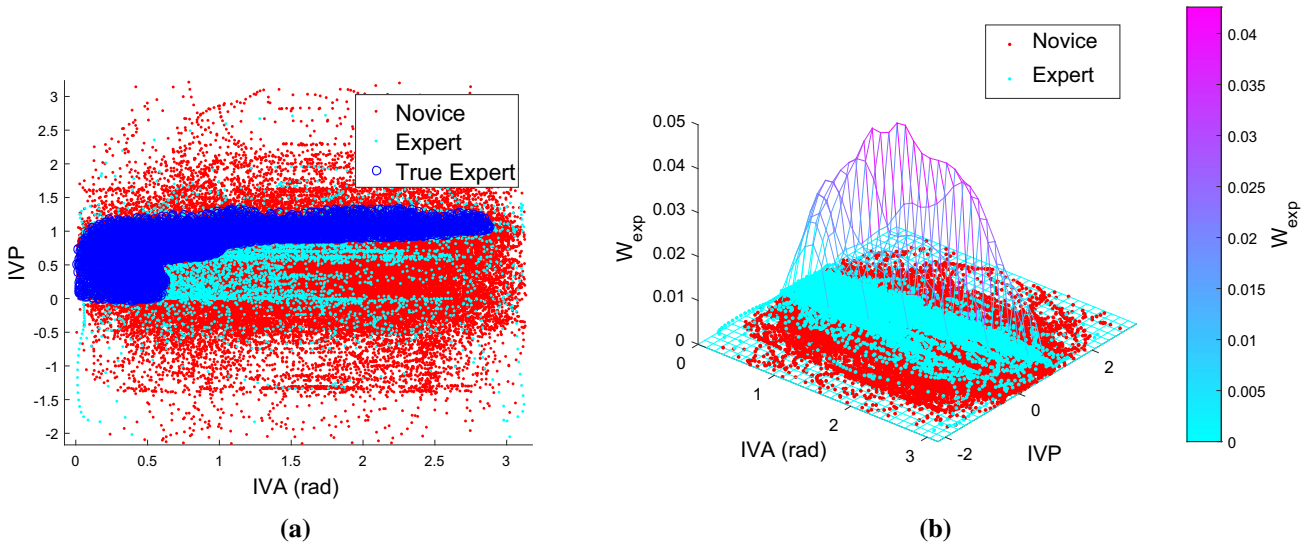

Fig. 5 Intent vector data with 'true expert' data and RELIEF-RBF weights (obvious novice and expert). **a** 'True expert' region. **b** RELIEF-RBF weights

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

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

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

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

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

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

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

[a] Achieves MAC criterion

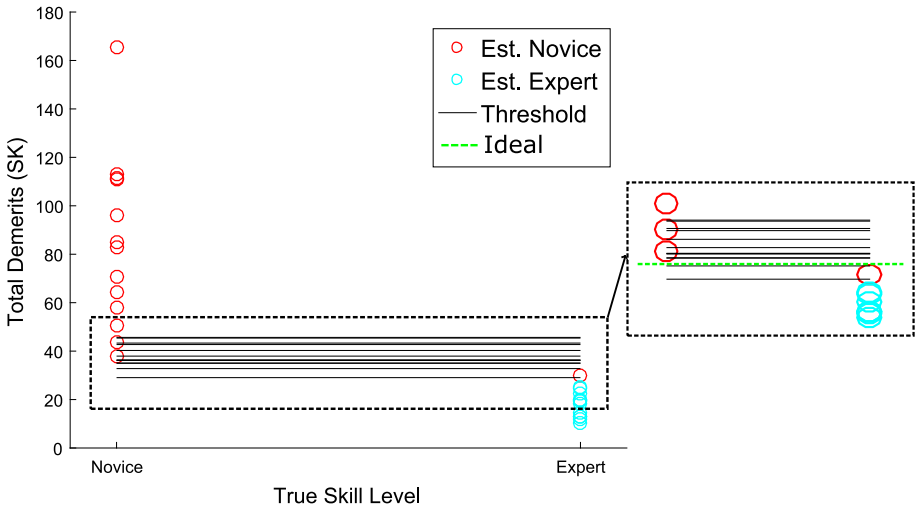

Fig. 6 LOUOpG classification using intent vectors with thresholds (T_{sk}) and ideal separable threshold

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

Conclusion

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

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

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

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

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

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

Compliance with ethical standards

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

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

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

References

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