Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks

Brian Allen · Vasile Nistor · Erik Dutson ·
Greg Carman · Catherine Lewis · Petros Faloutsos

Abstract

Background Despite technological advances in the tracking of surgical motions, automatic evaluation of laparoscopic skills remains remote. A new method is proposed that combines multiple discrete motion analysis metrics. This new method is compared with previously proposed metric combination methods and shown to provide greater ability for classifying novice and expert surgeons.

Methods For this study, 30 participants (four experts and 26 novices) performed 696 trials of three training tasks: peg transfer, pass rope, and cap needle. Instrument motions were recorded and reduced to four metrics. Three methods of combining metrics into a prediction of surgical competency (summed-ratios, z-score normalization, and support vector machine [SVM]) were compared. The comparison was based on the area under the receiver operating characteristic curve (AUC) and the predictive accuracy with a previously unseen validation data set.

Results For all three tasks, the SVM method was superior in terms of both AUC and predictive accuracy with the validation set. The SVM method resulted in AUCs of 0.968, 0.952, and 0.970 for the three tasks compared respectively with 0.958, 0.899, and 0.884 for the next best method (weighted z-normalization). The SVM method correctly predicted 93.7, 91.3, and 90.0% of the subjects' competencies, whereas the weighted z-normalization respectively predicted 86.6, 79.3, and 75.7% accurately ($p < 0.002$).

Conclusions The findings show that an SVM-based analysis provides more accurate predictions of competency at laparoscopic training tasks than previous analysis techniques. An SVM approach to competency evaluation should be considered for computerized laparoscopic performance evaluation systems.

Minimally invasive surgery (MIS) provides significant benefits to patients including shorter hospital stays, smaller scars, and faster healing. However, MIS procedures can be significantly more complex than their open procedure counterparts, and MIS thus requires longer training and additional experience.

Educational programs, such as the Fundamentals of Laparoscopic Surgery from the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) and the American College of Surgeons (ACS) are a significant step toward improved consistency and objectivity in surgical education, but many feel that further improvements in both quality and reduced training time are possible (see Aggarwal et al. [1] for a summary).

In recent years, technological advances in motion data acquisition for laparoscopic training such as virtual reality

B. Allen (✉) · P. Faloutsos
Department of Computer Science, University of California–Los Angeles, 4531-F Boelter Hall, Los Angeles, CA 90024, USA
e-mail: vector@cs.ucla.edu

V. Nistor · G. Carman
Department of Mechanical Engineering, University of California–Los Angeles, Los Angeles, CA, USA

E. Dutson · C. Lewis
Department of Surgery, University of California–Los Angeles, Los Angeles, CA, USA

(VR)-based [2, 3], optical (LapVR; Immersion Medical, 55 W. Watkins Mill Road, Gaithersburg, MD, USA), and magnetic [4] tracking systems have provided surgeons and their residents with copious data. Thus, although substantial kinematic data are available to judge the competency of a surgeon's performance, distilling useful automated feedback from this information remains difficult.

To meet this challenge, motion analysis systems generally reduce the full kinematic record to a small number of scalar metrics such as the time taken to complete the assigned task or the length of the path taken by the instrument tip over the course of a task. Studies have shown that principled combinations of metrics can provide a more powerful discriminator of competent versus noncompetent motions than a single metric [5, 6].

This report proposes a new approach for combining metrics based on the supervised machine-learning technique of support vector machines (SVMs) [7]. These SVMs provide a principled and automatic way to discover complex relationships between motion-derived metrics and the surgeon's level of prior training. The intuition that laparoscopic surgery is sufficiently difficult to render individual performance metrics nonlinearly interdependent provides the motivation to examine SVM-based approaches.

Methods

To evaluate the proposed technique, motion data were acquired using standard laparoscopic instruments in a training situation. These motion data were compiled into four scalar metrics, described in the Task Metrics section. The proposed SVM-based approach to combining these metrics (described in Combinations of Metrics section) was compared directly with the strongest previously reported methods: the summed ratios method [5] and the z-score normalization method [6]. Each of the three approaches was evaluated regarding its ability to predict the surgeon's prior level of experience based solely on these metrics computed from the motion data (see the Evaluating Classification Performance section).

Subjects

This study enrolled 30 individual participants, 4 of whom were practicing laparoscopic surgeons. The remaining 26 participants were medical school students and residents with no prior training in laparoscopic surgery. Each participant performed up to 10 trials on each of three evaluation tasks, for a total of 696 task performances across the two populations. Some participants conducted fewer than 30 trials because of time constraints.

Evaluation tasks

The surgical tasks recorded for this study were based on a set of previously validated laparoscopic training tasks [8–11]. The participant used a grasper instrument (Ethicon Endo-Surgery, Inc., 4545 Creek Road, Cincinnati, OH, USA) in the left hand and a needle driver (Karl Storz Endoscopy-America, Inc., 600 Corporate Pointe, Culver City, CA, USA) in the right hand, operating in a standard training box. Visual feedback was provided by a video camera at a resolution of 640×480 and a refresh rate of 30 Hz.

Using this setup, the participants were asked to perform 10 iterations each of three tasks (peg transfer, pass rope, and cap needle) in fixed order. The peg-transfer task required the participant to transfer a small ring made of rubber from a 1-in. peg to an identical peg several inches away by picking up the rubber piece with the instrument in the right hand, transferring it in midair to the instrument in the left hand, and then placing it on the peg. The participant then was required to transfer the rubber ring back to the original peg by reversing the order of the steps.

The pass-rope task required the participant to run a 10-in. cotton ''rope'' using a hand-to-hand technique. The rope was marked at 1-in. intervals indicating allowable grasp points. To evaluate each hand separately, the rope was first run toward the dominant hand and then toward the nondominant hand.

The cap-needle task required the participants to pick up a needle with their dominant hand and a needle cap with their other hand. Next, they fully inserted the needle into the cap without allowing either to touch the training box floor, and then placed the capped needle in a fixed position.

Motion-tracking system

Both of the laparoscopic instruments were modified to contain two electromagnetic sensors (Ascension Technology Corp., 107 Catamount Drive, Milton, VT, USA) capable of instantaneously reporting the location of the instruments. Placement of the sensors within the instruments is shown in Fig. 1. The sensors were sufficiently small (diameter, 1.3 mm) and lightweight (0.2 g, 11.8 g with cable) that the functionality of the instruments was not impaired. Each of the four sensors reported spatial position and orientation at 10 Hz with a linear accuracy of approximately 0.5 mm and an orientation accuracy of $0.2°$.

The position of the instrument's distal tip was calculated from the sensor information and recorded. The recorded motion of the instrument tip then was analyzed to generate the four metrics of task performance described in the next section.
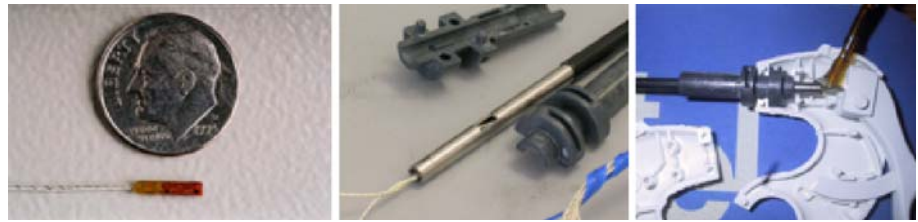
Task metrics

Four task-independent metrics $\{t_c, l, V c_e\}$ were gathered at each trial for each task. The first three metrics were kinematically derived from the motion data. The final metric, control effort, was estimated by analytical calculation of the forces applied to the instrument. All the metrics were scalar quantities computed by finite sums.

Each metric was appropriate to the simple training tasks observed and expected to have utility in measuring laparoscopic performance due to prior reports specific to laparoscopic surgery (time to completion and path length [12], control effort [6]), or extension from mentor-based skills assessments (volume [13]).

Time to completion (t_c) was the total time, measured in seconds, required by the participant to complete the assigned task and return the instrument tips to the starting positions. Path length (l) was the total linear distance measured in millimeters traveled by the distal tip of the instrument. Volume (V) was computed as the volume of the minimal axis-aligned bounding box that contained all samples of the distal sensor's position. Control effort (c_e) was a dynamic metric estimating the sum of forces rather than changes in position. This was possible because the tracking system recorded the orientation of the instruments as well as the position as a function of time.

This information provided accelerations, which, coupled with the measured inertial properties of the instrument, could be used to calculate forces. The calculation assumed the trocar was fixed in space, acting as an ideal, frictionless fulcrum. The masses of the peg from the peg-transfer task and the rope from the pass-rope task were assumed to be negligible. To estimate the control effort, the net force applied by the surgeon to the instrument was summed over the entire time to completion (t_c) of the task.

Combinations of metrics

In general, it was hypothesized that single metrics alone provided insufficient means to categorize the skill level with which a particular task was performed. In this section, we describe three methods that combine multiple metrics to label the subject with a single binary class, either "competent" (C +) or "noncompetent" (C −). The prior level of training of each participant (i.e., surgeon or resident)

was recorded and determined whether the participant was "expert" or "novice." Each of these three methods aimed to label experts automatically as competent (C +) and novices as noncompetent (C −) by examining only the motion metrics. The probability of a test reporting competency (C +) for an expert (E) was $P(C + |E)$ and known as the sensitivity (Sn) of the test. Conversely, the specificity of a test was the probability of reporting a novice as noncompetent: $Sp = P(C - |N)$.

The first two methods described later, summed ratios and z-score normalization, were derived from the literature and extended where needed. Both of these methods calculated an aggregate score (s), which then was compared with a cutoff score (s_c) such that competency was indicated by $s > s_c$. The third method was based on SVMs and did not require determination of a cutoff score.

Summed ratios

The summed ratios method [5] computes an aggregate score for an individual by summing normalized metrics. A metric is normalized by dividing the subject's metric (m) by the maximum score [$\max_E(m)$] obtained by an expert for that metric and associated task. All metrics are equally weighted:

$$s = \sum_{m \in \{t_c, l, V, c_e\}} \frac{m}{\max_E(m)}$$

The combined scores are classified into "competent" or "noncompetent" using a cutoff score that maximizes the product of sensitivity and specificity with equal weight: $s_c = \arg\max_s(Sp \times Sn)$.

Z-score normalization

The z-score normalization method [6] calculates the aggregate score (s) as the weighted average of z-scores obtained from metric values (m) using the mean [$\mu(m_E)$] and standard deviation (σ_E) of the expert data:

$$s = \sum_{m \in \{t_c, l, V, c_e\}} a_m \left(\frac{m - \mu(m_E)}{\sigma_E} \right),$$

where a_m is a scalar weight for metric m. We extend the treatment of Stylopoulos et al. [6] to find optimal values for the weights. This is done by considering the scalar weights as a single four-dimensional weight vector ($\vec{a}$). A series of

candidate weight vectors is generated with a per component step size of 0.1, and each is normalized, then evaluated as described in the Evaluating Classification Performance section. The weight vector with the best average performance is retained.

SVMs

In this section, we introduce a new method for classifying an individual's motion data based on SVMs. These SVMs offer a powerful method for automatically generating nonlinear functions from a set of labeled examples. One common use, and the one used in this study, is to generate functions that output a single binary datum, in this study, the competency with which a task is performed. More formally, for each individual in the training data, a vector is constructed containing a dimension for each explanatory variable.

In this study, each of the metrics was used as an explanatory dimension. A label, $z \in \{E, N\}$, was appended to store whether the measured motion was recorded from an expert or a novice to form the training vector $x = (t_c, l, V, c_e, z)$. Once the training process was completed, a new unlabeled vector of metrics (x') was given a label by determining the region where it fell $x' \mapsto \{C+, C-\}$. This label was a prediction of whether the subject was competent or noncompetent.

The SVM is trained by an iterative process of finding support vectors that divide the space of explanatory variables (in this case, the individual metrics) into expert and novice regions. Such support vectors are simply hyperplanes that separate training data points of different labels so that most expert points are on one side of the hyperplane and most novice points on the other side. Support vectors are chosen to maximize this separation of categories and to maximize the distance from the training points to the hyperplane itself. In this way, a small number of support vectors can efficiently partition the entire space of explanatory variables into separate regions, with each region related to one of the labels (i.e., E or N).

If the data are related in a linear manner, simpler methods, such as the z-score normalization described earlier, are sufficient. However, SVMs are able to handle nonlinear relationships between explanatory variables by using kernel functions [K(x,y)]. The kernel defines the distance function (i.e., the inner product) between two vectors of explanatory variables. A nonlinear kernel allows the linear separating hyperplanes to distinguish nonlinear relationships between the explanatory variables. The kernel can be understood intuitively as deforming the space containing the training points. When successful, this deformation permits the linear separating hyperplanes to account effectively for nonlinear relationships between the explanatory variables.

The implementation reported in this study uses libSVM [14], a freely available and open-source implementation of SVMs. Before training and classification, the input vector (x, z) is scaled linearly so that all elements are in [0,1], and the radial-basis function $K(x, y) = e^{\gamma \|x-y\|^2}$ is used as the kernel. The SVM training process uses a weighting factor C to scale the importance of errors in classifying the training data. The process of determining values for C and γ is described in the next section.

Evaluating classification performance

To compare the three methods of combining metrics described in the Combinations of Metrics section, we consider two approaches. The first is based on receiver operating characteristic (ROC) curve analysis and the second on validation against previously unseen data.

ROC analysis

The first approach relies on the ROC curve. The ROC curve is plotted as 1 minus the specificity versus the sensitivity. It provides an intuitive way to compare methods that accept the trade-off inherent in any binary classifier between being too sensitive and being too selective. The methods are compared quantitatively by the total area under their receiver operator characteristic curves (AUCs) using trapezoidal integration. Intuitively, the AUC estimates the probability that an expert chosen at random will score better than a randomly selected novice. Higher AUCs are more useful distinguishers, with an AUC of 1.0 being ideal and 0.5 no better than pure chance. The AUC is a common means of comparing diagnostic tests. Hanley and McNeil [15] show that the AUC is equivalent to the nonparametric Wilcoxon–Mann–Whitney statistic.

The second approach to comparing the three methods is to measure the accuracy of classification with a previously unseen data set. This validation process simulates the conditions of an online evaluation system deployed, for example, as a training assist to provide online, objective feedback. This approach is described in the next section.

Validation comparison

The motion data from each task are analyzed separately. The combined score for each trial is computed using one of the three methods of combination described previously. The following procedure is repeated 100 times for each method–task pair:

1. *Segment data.* The aggregate scores for each trial are randomly divided into two sets. Three-fourths of the scores are placed in a training set and the remaining

one-fourth in a validation set. Trials are drawn with uniform probability but adjusted as needed to preserve the approximate ratio of expert-to-novice trials in the generated sets.

2. *Determine method parameters using only the training set.*

   a. For the summed ratios, 500 candidate cutoff scores are tested across the full range of composite scores. The cutoff with the largest product of sensitivity and specificity is saved as the delineator.

   b. For the *z*-score normalization method, both a weight vector and its corresponding cutoff score must be determined. To find the best weight vector, each unique unit vector with elements in [0,1] and spaced by 0.1 is examined. For each of these candidate weight vectors, 500 candidate cutoff scores are tested over the range $[-2,2]$ (encompassing approximately 95% of the observed variance in composite scores). The combination of weight vector and cutoff score that produces the largest product of sensitivity and specificity is saved as the delineator.

   c. For the SVM method, C and γ are found using an exponential grid search across $C = 2^{17}, 2^{15}, \ldots, 2^{-3}$ and $\gamma = 2^{-17}, 2^{-15}, \ldots, 2^{3}$. Each (C, γ) pair is evaluated by performing fivefold cross-validation using only the training set data. The support vectors producing the highest accuracy rate are used for the classifier.

3. *Evaluate against the validation set.* The validation set then is classified using the parameters determined in training (i.e., cutoff score, weight vector and cutoff score, or support vectors). The resulting accuracy, specificity, and sensitivity of classification are computed over the validation set.

The significance of the validation set accuracy is calculated using Welch's *t*-test [16], and a threshold of 0.05 is considered significant.

Results

Individual metrics

Table 1 summarizes the values calculated for each individual metric for each task: peg transfer ($n = 285$: 31 expert and 254 novice), pass rope ($n = 212$: 29 expert and 183 novice), cap needle ($n = 199$: 30 expert and 169 novice). All four individual metrics were computed for each subject's attempt at each task. Table 1 provides the

Table 1 Mean ± standard deviation values for each of the individual metrics applied to each task and for each population subgroup

	All	Expert	Novice
Path length (mm)			
Peg transfer	1316.5 ± 719.1	701.8 ± 138.3	1391.5 ± 725.5
Pass rope	2141.1 ± 904.1	1464.8 ± 350.3	2248.3 ± 918.9
Cap needle	1456 3 ± 920.9	907.2 ± 253.1	1553.8 ± 961.8
Time (s)			
Peg transfer	45.44 ± 24.54	24.41 ± 5.22	48.01 ± 24.74
Pass rope	92.18 ± 31.62	61.77 ± 9.59	97.00 ± 31.22
Cap needle	50.99 ± 34.93	26.07 ± 8.98	55.41 ± 35.96
Volume (cm^3)			
Peg transfer	2.559 ± 2.442	1.279 ± 0.392	2.715 ± 2.539
Pass rope	1.634 ± 2.176	0.923 ± 0.842	1.753 ± 2.300
Cap needle	5.025 ± 3.025	4.133 ± 1.350	5.184 ± 3.210
Control effort (m^2/s^2)			
Peg transfer	10.51 ± 16.71	3.58 ± 2.08	11.36 ± 17.50
Pass rope	15.07 ± 12.47	9.62 ± 6.77	15.93 ± 12.95
Cap needle	12.08 ± 12.99	6.63 ± 4.17	13.04 ± 13.78

mean and standard deviation of the collected metrics for the different populations (i.e., novices, experts, and all participants).

Figure 2 presents histograms for the data summarized in Table 1. Each combination of task and metric is shown as two superimposed histograms. The two histograms are derived from disjoint distributions: one for the expert performances and one for the novice performances. The vertical axis measures the number of performances in each bin, with metrics in the range indicated on the horizontal axis. The vertical dotted line in each graph indicates the score that optimally separates the novice and expert populations. This separating score is determined by maximizing the product of specification and sensitivity for the entire sample.

The histograms illustrate that for each metric–task pair studied, the optimal separating line fails to divide cleanly novice from expert. Although the distributions are qualitatively different, the significant overlap reduces the usefulness of the metric for distinguishing novice from expert.

It should be noted also that the separating power of a single metric varies with the task. For example, control effort has little ability to distinguish between the two sample groups for the cap-needle and pass-rope tasks, but it is quite effective for the peg-transfer task.

Combinations of metrics

The methods for combining metrics are first compared by AUC. Total AUC measurements for each task and each method are provided in Table 2. Values closer to 1 indicate

Fig. 2 Four histograms comparing the frequency distributions of observed metrics for the peg-transfer task (**A**), the pass-rope task (**B**), and the cap-needle task (**C**). The darker regions show the experts' score distributions, and the lighter regions show those of the novices. The dotted vertical line indicates the optimal separating score
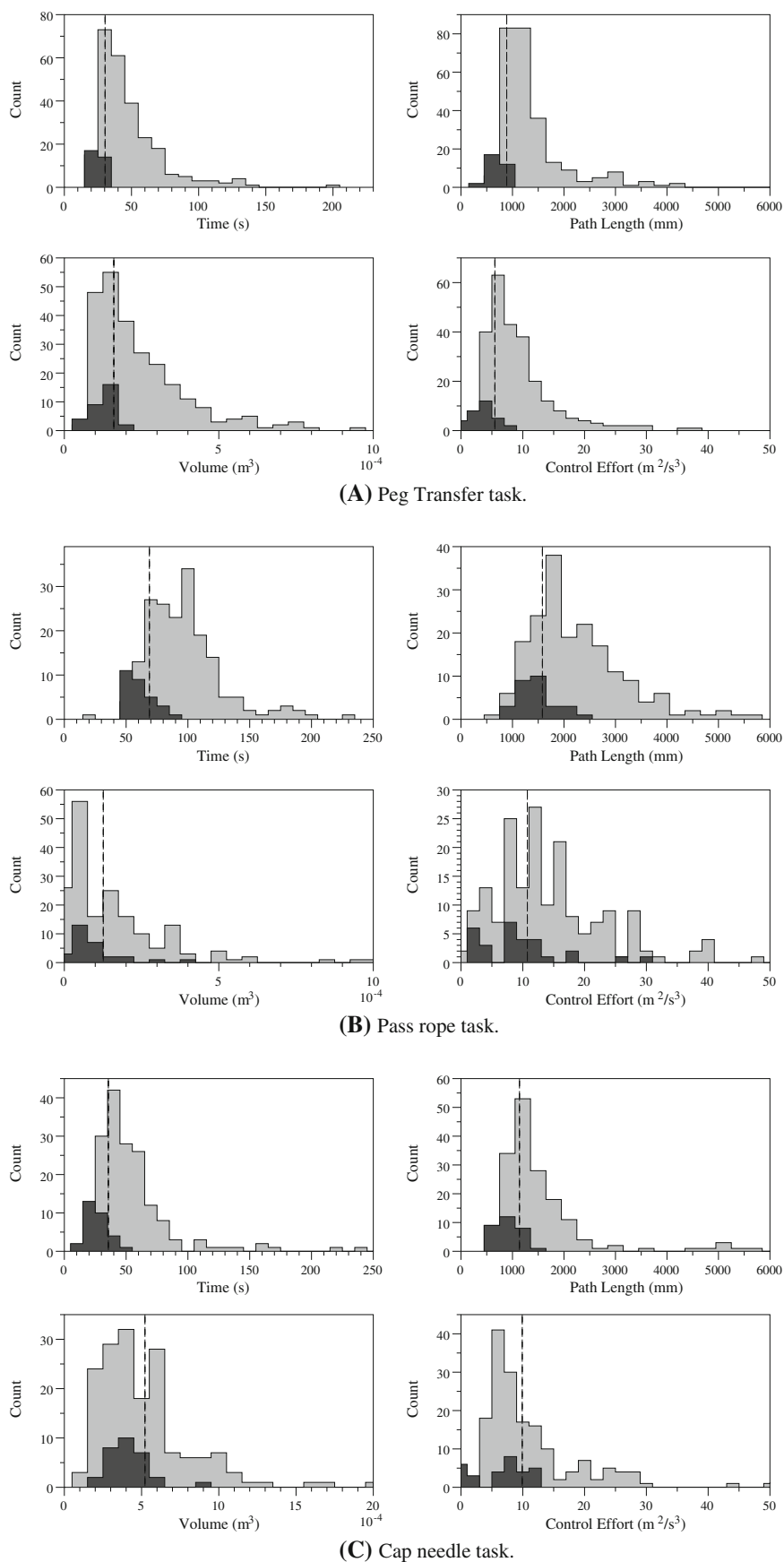

(A) Peg Transfer task.

(B) Pass rope task.

(C) Cap needle task.

Table 2 Comparison of area under receiver operating characteristic (ROC) curves (AUC) for each method of combining individual metrics

Method	Peg transfer	Pass rope	Cap needle
SVM	0.9682	0.9520	0.9704
Weighted *z*-normalization	0.9582	0.8994	0.8840
Summed ratios	0.9444	0.8356	0.7834

SVM support vector machine

better ability to distinguish novice from expert. The three methods of metric combination are ordered consistently by AUC over all three tasks. The SVM method outperforms the weighted *z*-normalized method, which in turn outperforms the summed-ratios method.

The SVM method is consistently the best classifier, as measured by AUC. The cap-needle task shows the largest difference, with an area of 0.9704 for the SVM method compared with 0.884 for the weighted *z*-normalized method or 0.7834 for the summed-ratios method. The *z*-normalized method has the largest mean AUC using a mean best-performing weight vector of 0.965, 0.0, 0.033, or 0.002. The best weight vectors for the peg-transfer and pass-rope tasks are, respectively, 0.034, 0.865, 0.005, 0.096 and 0.883, 0.086, 0.0, 0.031.

The second approach to comparing the methods for combination of metrics is by their accuracy in classifying previously unseen data, which shows the SVM-based method to be more accurate and to yield a smaller variance. Table 3 shows the mean accuracy and standard deviation for each method over the data for each training task. The accuracies of the SVM predictions are significantly better

Table 3 Mean accuracy [$\mu(Acc)$] and standard deviation of the accuracy [$\sigma(Acc)$] for each task and method pair are calculated for more than 100 repetitions of classification on a randomly selected validation set using the mean cutoff score [$\mu(s_c)$] as the mean score to divide predicted expert from novice

Method	$\mu(Acc)$ (%)	$\sigma(Acc)$ (%)	$\mu(s_c)$ (%)
Peg transfer			
SVM	93.7	2.6	NA
Weighted *z*-normalization	86.6	7.0	1.405
Summed ratios	83.2	5.1	3.783
Pass rope			
SVM	91.3	4.3	NA
Weighted *z*-normalization	79.3	9.6	1.071
Summed ratios	72.2	7.3	3.454
Cap needle			
SVM	90.0	3.4	NA
Weighted *z*-normalization	75.7	13.5	0.803
Summed ratios	70.8	6.9	3.658

SVM support vector machine, *NA* not available

than the next best method (weighted *z*-normalized) for all three training tasks: peg transfer ($p < 0.001$), pass rope ($p = 0.001$), and cap needle ($p = 0.002$).

Figure 3 provides a graphic comparison of the three methods for combining metrics by showing the ROC curves for each task. For all three tasks, the SVM method dominates both alternatives. That is, for each given sensitivity, the SVM method provides equal or higher specificity, resulting in a curve above and to the left of the other curves. Figure 3 also suggests that as task complexity increases, from the simplest task (peg transfer) to the most complex (cap needle), motion analysis metrics become less useful in distinguishing different competencies. However, the SVM method shows markedly less performance degradation than either the weighted *z*-normalized or summed-ratios methods.

Discussion

Our results show that the accuracy of competency prediction can be dramatically improved (by 7, 12, and 14% for the tasks examined, Table 3) simply by improving the analysis of the motion data. Because this finding builds on a standard motion tracking approach, it is likely robust to differences in specific technology and platform and thus widely applicable. To our knowledge, this study provides the first direct comparison of aggregation techniques applied to the analysis of laparoscopic motions.

It is important to note that our method does not rely merely on linear relationships between metrics, so even if a given metric is poorly correlated with competent performance overall, it may add to the analysis as a whole. The recent study by Chmarra et al. [17] shows that minimization of the path length likely is not characteristic of expert surgeons. Reporting merely the raw metric data or a linear combination of such to a student is unlikely to provide the ideal feedback. It is even possible that presenting overly simplified metrics, such as path length or time to completion, directly to the student will encourage the student to maximize those metrics at the expense of overall competency.

The characteristics of SVMs as an analysis tool are well matched to the problem of judging surgical competence based on motion data. First, because the SVM learns from example motions, the effectiveness of an SVM-based performance evaluator stems from actual differences in the motions of experts and novices. This can be contrasted with attempts to determine artificially the quality or importance of individual metrics. Second, SVM classifiers are able to integrate several orders of magnitude more example motions than used in this study while still providing rapid responses to new queries [7]. Third, as new metrics are
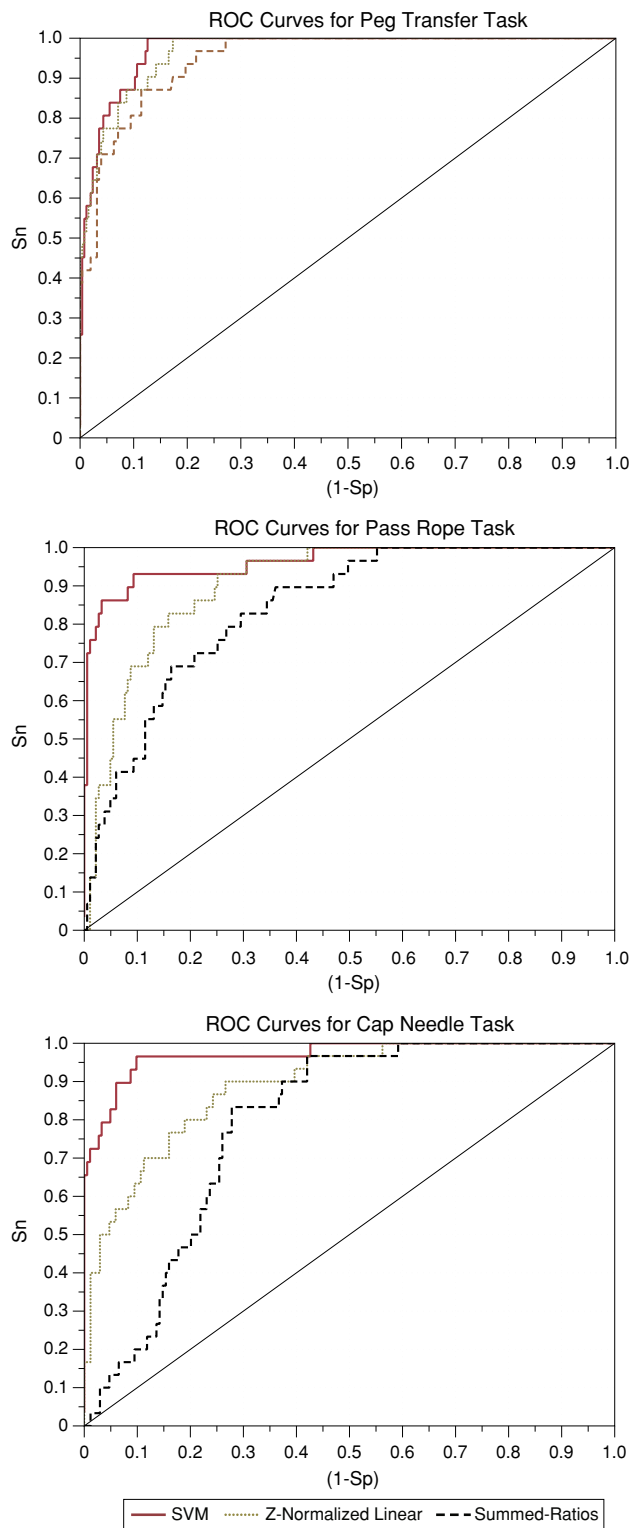
Fig. 3 Receiver operating characteristic (ROC) curves for all three tasks: peg transfer (*top*), pass rope (*center*), and cap needle (*bottom*). The straight diagonal line represents a theoretical random classifier

devised, they can be added trivially to the evaluator to improve accuracy. To our knowledge, this study is the first to suggest the use of SVMs in this domain, although they

are becoming a common approach to a variety of difficult diagnostic problems.

Our analysis considers aggregate kinematic information and is capable of evaluating only low-level motor skills. Higher-level surgical skills are not examined, and approaches based on aggregate metrics are unlikely to provide any significant insight because their aggregate nature obfuscates the strategies and intentions of the surgeon. However, our findings also show that the simplest laparoscopic training task (peg transfer) benefits least from the proposed use of SVMs. Perhaps this is because some of the novice participants who previously attained a sufficient level of competency at this task were indistinguishable from the experts. In effect, a few novices were already experts at the peg-transfer task.

Such acquisition of rudimentary skills through nonsurgical activities, such as video games, has been documented previously [18]. If this is the case, then the described approach is likely to be most useful for evaluating tasks of intermediate complexity: complex enough to require significant motor skills yet simple enough not to require high-level or strategic surgical abilities. Figure 3 suggests that motion analysis methods in general have difficulty as task complexity increases. However, the SVM method proves more robust to increased complexity, suggesting that it may be useful for even more complex motor skills, such as knot tying.

A trade-off to the power of the SVM to model nonlinear relationships between metrics is that the resulting support vectors can be difficult to understand intuitively. Thus, it may be difficult to explain to a student precisely why his or her performance was classified as it was.

In conclusion, the maturation of laparoscopic training systems is providing a wealth of data tracking of trainees' movements. New techniques are needed to take full advantage of the ability these systems have to evaluate surgical performance. This study demonstrates that improved analysis of motion data can increase the accuracy and discriminatory power of existing and future computer-enhanced training systems.

References

1. Aggarwal R, Moorthy K, Darzi A (2004) Laparoscopic skills training and assessment. Br J Surg 91:1549–1558
2. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, Satava RM (2002) Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 236:458–463 (discussion 463–464)

3. Maithel S, Sierra R, Korndorffer J, Neumann P (2006) Construct and face validity of MIST-VR, Endotower, and CELTS: are we ready for skills assessment using simulators? Surg Endosc 20:104–112

4. Nistor V, Allen B, Carman G, Faloutsos P, Dutson E (2007) Haptic-guided telementoring and videoconferencing system for laparoscopic surgery. In: Proceedings of the SPIE 14th international symposium on smart structures and materials and nondestructive evaluation and health monitoring, San Diego, CA, 18–23 March 2007

5. Fraser S, Klassen D, Feldman L, Ghitulescu G (2003) Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. Surg Endosc 17:964–967

6. Stylopoulos N, Cotin S, Maithel SK, Ottensmeyer M, Jackson PG, Bardsley RS, Neumann PF, Rattner DW, Dawson SL (2004) Computer-enhanced laparoscopic training system (CELTS): bridging the gap. Surg Endosc 18:782–789

7. Vapnik VN (1998) Statistical learning theory. Wiley, New York

8. Donias HW, Karamanoukian RL, Glick PL, Bergsland J, Karamanoukian HL (2002) Survey of resident training in robotic surgery. Am Surg 68:177–181

9. Nio D, Bemelman WA, Boer KT, Dunker MS, Gouma DJ, Gulik TM (2002) Efficiency of manual versus robotical (Zeus) assisted laparoscopic surgery in the performance of standardized tasks. Surg Endosc 16:412–415

10. Garcia-Ruiz A, Gagner M, Miller JH, Steiner CP, Hahn JF (1998) Manual vs robotically assisted laparoscopic surgery in the performance of basic manipulation and suturing tasks. Arch Surg 133:957–961

11. Peters JH, Fried GM, Swanstrom LL, Soper NJ, Sillin LF, Schirmer B, Hoffman K, The SAGES FLS Committee (2004) Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 135:21–27

12. Aggarwal R, Grantcharov T, Eriksen J, Blirup D (2006) An evidence-based virtual reality training program for novice laparoscopic surgeons. Ann Surg 244:310–314

13. Adrales G, Park A, Chu U, Witzke D (2003) A valid method of laparoscopic simulation training and competence assessment. J Surg Res 114:156–162

14. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/. Accessed 15 Oct 2008

15. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29

16. Welch B (1947) The generalization of "student's" problem when several different population variances are involved. Biometrika 34:28–35

17. Chmarra MK, Jansen FW, Grimbergen CA, Dankelman J (2008) Retracting and seeking movements during laparoscopic goal-oriented movements: is the shortest path length optimal? Surg Endosc 22:943–949

18. Rosser J, Lynch P, Cuddihy L, Gentile D, Klonsky J, Merrell R (2007) The impact of video games on training surgeons in the 21st century. Arch Surg 142:181–186