RESEARCH ARTICLE JSLS

Surgical Performance Analysis and Classification Based on Video Annotation of Laparoscopic Tasks

Constantinos Loukas, PhD, Athanasios Gazis, MSc, Meletios A. Kanakis, MD

ABSTRACT

Background and Objectives: Current approaches in surgical skills assessment employ virtual reality simulators, motion sensors, and task-specific checklists. Although accurate, these methods may be complex in the interpretation of the generated measures of performance. The aim of this study is to propose an alternative methodology for skills assessment and classification, based on video annotation of laparoscopic tasks.

Methods: Two groups of 32 trainees (students and residents) performed two laparoscopic tasks: peg transfer (PT) and knot tying (KT). Each task was annotated via a video analysis software based on a vocabulary of eight surgical gestures (surgemes) that denote the elementary gestures required to perform a task. The extracted metrics included duration/counts of each surgeme, penalty events, and counts of sequential surgemes (transitions). Our analysis focused on trainees' skill level comparison and classification using a nearest neighbor approach. The classification was assessed via accuracy, sensitivity, and specificity.

Results: For PT, almost all metrics showed significant performance difference between the two groups ($p < 0.001$). Residents were able to complete the task with fewer, shorter surgemes and fewer penalty events. Moreover, residents performed significantly fewer transitions ($p < 0.05$). For KT, residents performed two surgemes in significantly shorter time ($p < 0.05$). The metrics derived from the video annotations were also able to recognize the trainees' skill level with 0.71 – 0.86 accuracy, 0.80 – 1.00 sensitivity, and 0.60 – 0.80 specificity.

Conclusion: The proposed technique provides a tool for skills assessment and experience classification of surgical trainees, as well as an intuitive way for describing what and how surgemes are performed.

Key Words: Video Analysis, Video Annotation, Laparoscopic Training, Skills Assessment, Classification.

INTRODUCTION

Surgical skills assessment has been a major field of research for many years. Traditionally, the assessment is based on the teacher-apprentice model where a faculty surgeon evaluates the student during task performance. However, this type of evaluation is known to be biased, subjective, qualitative, and limited in terms of its ability to provide constructive feedback to the trainee.[1] Several studies have outlined the need for objective and quantitative measures of performance during training.[2,3] The assessment should be designed so that it provides specific, rather than abstract, metrics of performance in order to allow trainees to have an understanding of what should have been done or avoided.

A common way for assessment of technical skills is based on task-specific checklists and global rating scales, such as the objective structured assessment of technical skill (OSATS) and global operative assessment of laparoscopic skills (GOALS) tools.[4,5] The assessment is typically performed by a faculty surgeon after reviewing the video of the task/procedure performed, a process that is time consuming, perceptual demanding, and prone to inter-observer variability. Recent changes in accreditation and residency training programs require continuous evaluation and development of surgical skills, highlighting the need for objective assessment measures of task performance and of the fine-grained actions performed.[6,7]

An alternative approach to skills assessment is based on motion analysis of surgical gestures via specialized sensors (electromagnetic, infrared, etc.), attached to the

Medical School, National and Kapodistrian University of Athens, Athens, Greece (Drs. Loukas and Gazis).

Department of Pediatric and Congenital Heart Surgery, Onassis Heart Surgery Centre, Athens, Greece. (Dr. Kanakis)

Disclosure: none.

Funding/Financial support: none.

Conflicts of Interest: The authors declare no conflict of interest.

Informed consent: Dr. Constantinos Loukas declares that written informed consent was obtained from the patient/s for publication of this study/report and any accompanying images.

Address correspondence to: Dr. Constantinos Loukas, Medical School, National and Kapodistrian University of Athens, Mikras Asias 75 str., Athens 11527, Greece. Telephone: +30-210-7462437; Fax: +30-210-7462369, E-mail: cloukas@med.uoa.gr.

DOI: 10.4293/JSLS.2020.00057

surgeon's hands or instruments. A plethora of systems have been developed over the past few years,[8] showing strong correlation between expertise level and motion parameters such as instrument path length, number of movements, and advanced metrics based on computational models.[9–11] Although accurate in surgical level recognition, these sensor-based systems exhibit some limitations such as high investment cost, potential interference in task performance, and modification of the training setup. Moreover, although the extracted metrics may lead to greater accuracy, sometimes they cannot provide meaningful feedback to the trainee due to their abstract interpretation.[12] Hence, potential deficiencies in certain surgical skills cannot be always addressed adequately.

Video-based skills assessment has been introduced as an alternative approach to alleviate the limitations of sensor-based systems. In particular, the video signal from the endoscopic camera provides a direct information source for skills assessment, without the requirement for employing additional sensors. The main challenge lies in the extraction of visual features that capture not only the skill level of the trainee, but also the underlying relation to more semantic measures of performance, such as the OSATS criteria.[13] Various skill assessment approaches have been proposed in the literature based on features extracted from still frames,[14,15] video sequences,[16,17] or combination of video and motion data.18,[19] However, as described above, the measures extracted are often abstract and thus lack educational interpretation. Recent studies have attempted to address this issue by correlating the video-based performance metrics with the OSATS criteria used in surgical training.[20,21]

The aim of this study is to present a proof-of-concept methodology for performance comparison and experience recognition of surgical trainees, based on video annotation of surgical training tasks. The annotation was based on a vocabulary of surgical gestures, specifically developed for two basic laparoscopic tasks (peg transfer and knot tying). By means of this vocabulary and a free video annotation software, we extracted various interpretable measures of surgical performance. The proposed methodology was evaluated for its potential: (a) to assess differences in task performance among individuals with variable surgical experience, (b) to provide a tool for skills recognition, and (c) to provide an easy-to-use platform for skills assessment. Compared to previous approaches, the proposed method provides outcome measures with meaningful interpretation of surgical performance and yet does not require additional installation, cost, or surgical expertise to perform the video annotation.

MATERIALS AND METHODS

Subjects and Equipment

Seventeen year 2–3 medical students (MS), and 15 post-graduate year 1–3 surgical residents (RS), participated in the study. Prior to enrollment a written informed consent was obtained from all participants. The participants also completed a questionnaire about their demographic information and experience in surgery and video games (see **Table 1**). The MS group had no prior experience in laparoscopic training whereas the residents had assisted in or performed 12 [0–17] (median [range]) laparoscopic cholecystectomy operations.

The laparoscopic equipment included: a box trainer, surgical training models (pegboard, pegs, tissue pad, and suture with needle), two graspers, a rigid endoscope connected to the video processing unit of a laparoscopic tower, an LCD monitor, a custom-made tripod mount for the endoscope, and a DVD recorder for video task recording. The study initially included an instructional phase during which the subjects received a tutorial about the equipment and the tasks that had to be performed. The subjects were also allowed to perform a familiarization trial for each task.

Task Description

In the next phase, each subject performed two basic laparoscopic tasks: peg transfer (PT) and knot tying (KT). The goal of PT was to place four cylindrical pegs into the holes of a pegboard. The pegs were placed on either side (left or right) of the cavity of the box trainer. For the first two pegs, the subject had to use the left/right grasper to place a peg into a hole located at the same side of the pegboard.

Table 1.
Demographic Data per Group

	Medical Students	Surgical Residents
Number of participants	17	15
Age (average)	20.3	31.6
Sex, (M/F)	10:7	2:1
Hand dominance (L/R)	0:17	0:15
Surgical experience (observed, Y/N)	4:13	15:0
Surgical experience (assisted or performed, Y/N)	0:17	15:0
Video game experience (Y/N)	5:12	4:11

For the other two pegs the user had to transfer the peg on to the other grasper in order to place the peg on that side of the pegboard. **Figure 1** shows sample video frames of the main gestures performed: reach for peg, place peg, and peg transfer.

For KT, the goal was to perform a single loop knot by manipulating a suture with a needle attached at one end. The suture was predriven through a tissue pad. To complete the task, the subject had to pick up and orient the needle with one grasper, make a C loop, and finally reach and pull the free end of the suture with the other grasper. **Figure 2** shows sample video frames for the main gestures performed during the KT task: reach for needle, orient needle, make a C-loop, reach for suture, and pull suture.

Surgical Motion Vocabulary and Annotation

The videos of the training tasks were annotated for elementary gestures using the free video annotation software Anvil 6.0.[22] The annotation was based on the language of surgery concept, according to which a surgical motion is a composition of elementary activities that are sequentially performed with certain constraints.[23] The language of surgical motion focuses on the description of specific actions (surgemes) that surgeons perform with the instruments to achieve an intended goal. The surgemes define the elementary surgical gestures performed to complete the task. In this study we defined a vocabulary of eight surgemes to describe the main activities involved in the performance of the two tasks (three surgemes for PT and five for KT). Consequently, each task performance was represented by a sequence of surgemes, where each surgeme corresponded to a time interval with an absolute start and end time. Example images of the surgemes defined for each task are given in **Figure 1** (PT task) and **Figure 2** (KT task).

We also defined two penalty events (one per task), which correspond to a single time point event: peg drop and needle drop. These events represented the unintended result of a surgeme. Based on this definition, an event signified the (unintended) end of a gesture and the beginning of a new one. For example, in the PT task, the user may accidentally drop a peg while attempting to place it on the pegboard. In this case, the event: 'peg is dropped' signifies the end of the 'place peg' surgeme and the beginning of a new one ('reach for peg'). A description of the surgemes/events employed in the annotation of the videos is given in **Table 2**. Based on the previous description, an ideal task performance (i.e. without penalty events and unnecessary gestures) consists of the following sequential surgemes:

- PT task: RP → PP → RP → PP → RP → TP → PP → RP → TP → PP
- KT task: RN → ON → CL → RS → PS

Performance Metrics and Statistics

The video annotation was performed by an individual in consultation with a surgeon, based on the information

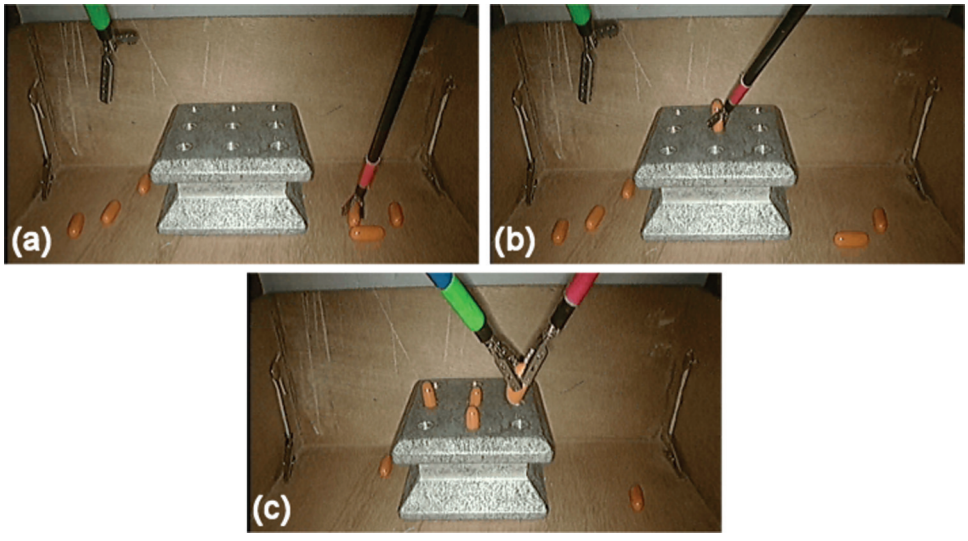

Figure 1. Sample images showing the three main gestures performed during the peg transfer task.

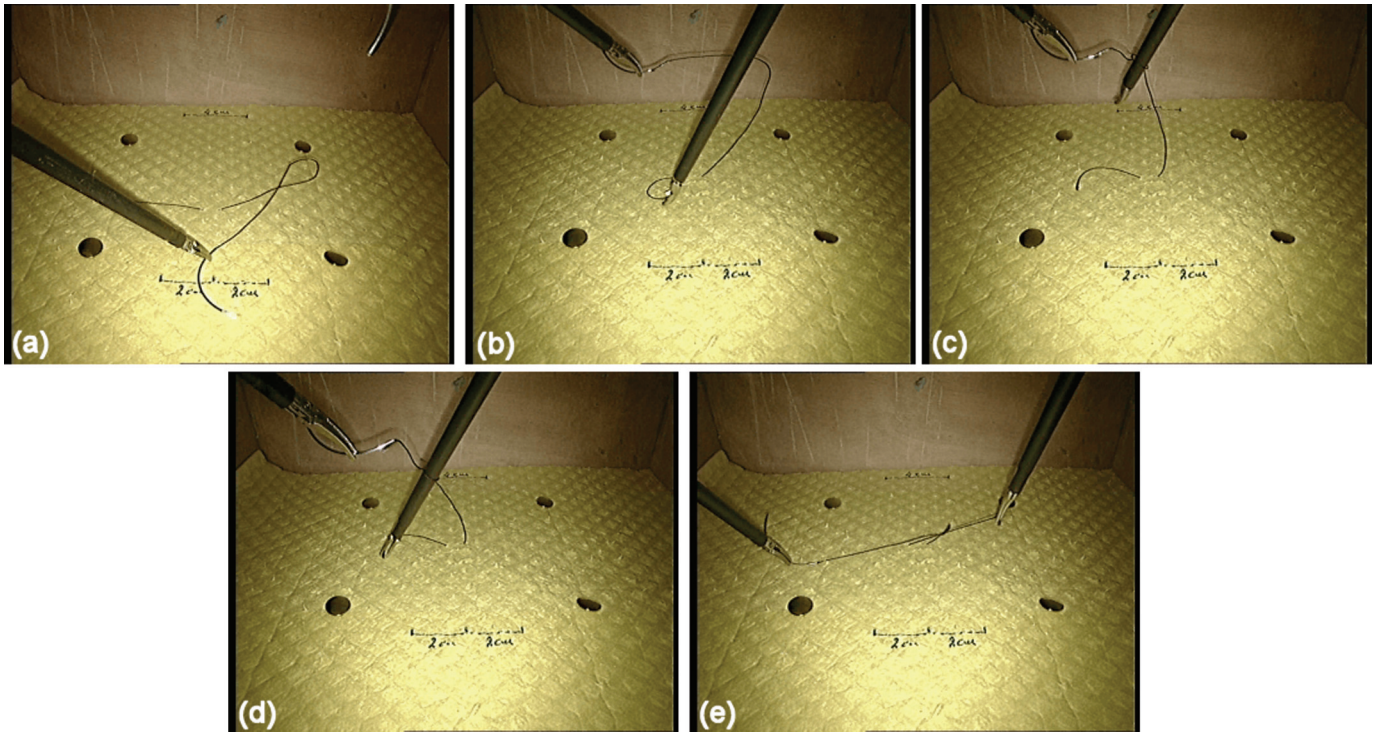

Figure 2. Sample images showing the five main gestures performed during the knot tying task.

shown in **Table 2**. The annotation output was an xml file (one file per task and subject), with the start/end time of the surgemes performed, and the timings of potential penalty events. From this information the following metrics were extracted: surgeme counts, duration of each surgeme, counts of penalty-events, and counts of surgeme transitions (e.g. RP→PP, RN→ON, etc.). The later metric allowed the construction of an $N \times N$ matrix, where N is the number of surgemes defined per task ($N_{PT} = 3$ and $N_{KT} = 5$). Some surgeme pairs were not observed/valid, so the total number of allowable transitions was: $m_{PT} = 6$ (PP→RP, RP→PP, RP→RP, RP→TP, TP→PP, TP→RP) and $m_{KT} = 7$ (CL→RN, CL→RS, ON→CL, ON→RN, RN→ON, RN→RN, RS→PS). Invalid transitions denote transitions that are not possible to occur. For example, in the PT task a new peg cannot be transferred right after a peg is placed into the pegboard (i.e. transition PT- ≥ TP is invalid). The trainee has first to reach for a new peg and then transfer it into the other grasper (transition RP- ≥ TP). Alternatively, depending on the step that is currently performed, the new peg may be placed into the pegboard (transition RP- ≥ PP).

Table 2.
Surgical Gestures and Events Annotated in the Task Performance Videos

Surgeme/Event Prefix	Task	Description
RP	PT	Reach for peg.
TP	PT	Transfer peg between graspers.
PP	PT	Place peg into pegboard.
RN	KT	Reach for needle.
ON	KT	Orient needle.
CL	KT	Making C-loop around right grasper.
RS	KT	Reach for suture with right grasper.
PS	KT	Pull suture with both graspers.
DP	PT	Peg is dropped (event).
DN	KT	Needle is dropped (event).

Recognition of Experience Level

In addition to the comparison of the two groups, we also examined whether the derived metrics are able to recognize the experience level of each participant (student or resident).

To ensure validity of the experience groups, we asked two expert surgeons (E1 and E2) to independently review and categorize the videos into the two groups. Both raters showed strong agreement with respect to the self-proclaimed skill labels (E1: agreement = 0.937; Cohen's $\kappa = 0.874$, $P \leq .01$; E2: agreement = 0.906; Cohen's $\kappa = 0.811$, $P \leq .01$).

Given a certain metric and the known experience level of the participants, a k-nearest neighbor (kNN) approach was employed. In particular, the (unknown) experience level of a candidate participant was determined from the Euclidean distances between his/her metric value and the metric values of the other participants (ground-truth). Hence, the experience level (student or resident) with the highest votes among the kNN was assigned to the candidate participant. Various values for k were examined: $k = \{1,3,5,7\}$. The kNN approach was followed separately for each task and metric in order to evaluate the best combination. The following evaluation metrics were employed to measure the method's performance:

$$Accuracy = \frac{TP+TN}{P+N} \quad (1)$$

$$Sensitivity = \frac{TP}{P} \quad (2)$$

$$Specificity = \frac{TN}{N} \quad (3)$$

where TP, TN, FP, FN, P, N denote: true positives, true negatives, false positives, false negatives, positives and negatives, respectively ($P = TP + FN$, $N = TN + FP$). The class definition was based on the convention: SR = positive class and MS = negative class. Hence, higher Sensitivity/Specificity denote better recognition of the Residents/Students class respectively.

The MATLAB® Statistics toolbox ver. R2018a (MathWorks, Natick, MA, USA) was used for data analysis. The performance metrics of the two groups were compared with the Mann-Whitney U test. A p-value of less than 0.05 was considered statistically significant. Data are presented as: medians [25th quartile – 75th quartile], unless otherwise specified.

RESULTS

Figures 3 and **4** show video annotation examples from PT and KT, respectively. The right window displays the sequential video frames whereas the left windows display information about video metadata and the defined coding scheme. The bottom window shows a color-coded timeline of the annotated surgemes (top row) and time-point events (bottom row). For better visualization, only a small segment of the entire timeline is shown. As shown in the Figures, a 'drop event' occurred when the user attempted to place the peg on the pegboard (PT task, **Figure 3**) and to orient the needle (KT task, **Figure 4**). These events signify an unintended

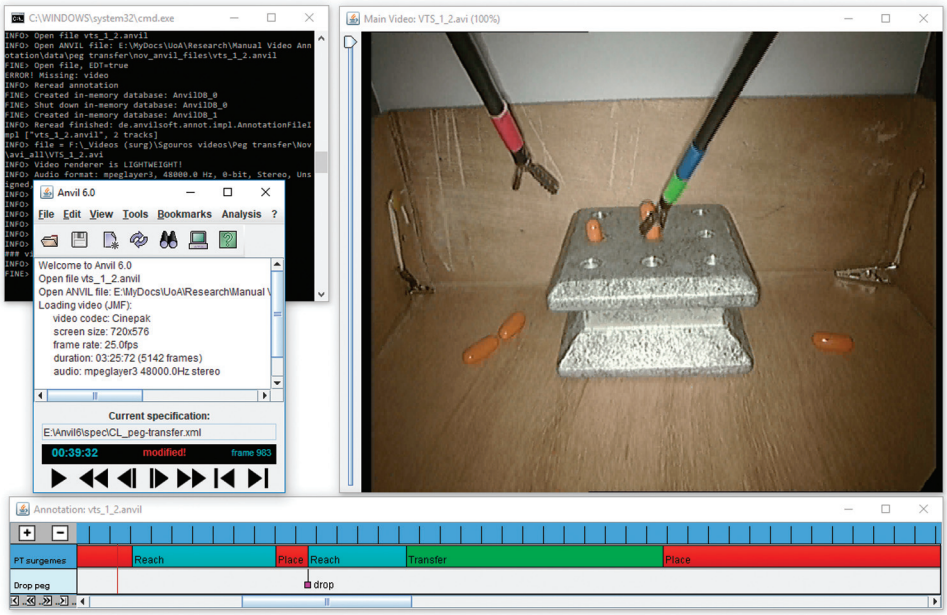

Figure 3. Snapshot from the annotation of a peg transfer video performance.

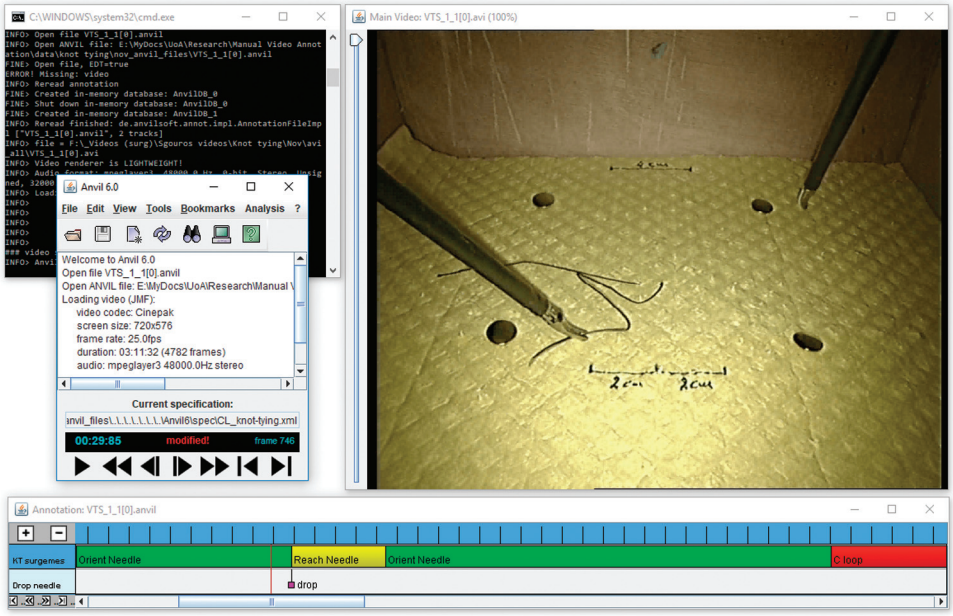

Figure 4. Snapshot from the annotation of a knot tying video performance.

transition between two surgemes (place peg and reach peg; orient needle and reach needle).

Tables 3 and **4** show median values and interquartile range for the metrics extracted from the PT and KT videos respectively, as well as statistical comparison between the groups. For PT, almost every metric showed a highly significant difference between the groups ($P < .001$). RS were able to complete the task with fewer and shorter surgemes as well as fewer penalty events compared to MS. Moreover, the interquartile difference of RS is clearly smaller than that of the MS across all metrics, denoting better performance. The only metric that did now show significance was the TP counts (i.e. peg transfers between the graspers), probably due to the predetermined number of transfers required by the task (two transfers).

For KT, only two metrics showed significant difference between the groups: RN and ON duration ($P < .05$). The residents required significantly less time than students to pick up and orient the needle. For the other metrics, the residents seem to perform fewer and shorter surgemes as well as fewer errors than students, but this trend was not significant ($P > .05$).

Figures 5 and **6** show statistical comparison results (p values) with respect to the surgeme transitions during PT and KT, respectively. The gray-shaded boxes indicate not allowable/observed transitions. Next to the table is the corresponding surgeme transition diagram, where significant transitions are drawn with increased weight. For the PT task, residents performed significantly fewer PP→RP (3 vs. 6), and RP→PP (2 vs. 4) transitions compared to students ($P < .05$). Retrospective evaluation of the videos showed that this finding was due to the 'drop peg' counts, which occurred mostly during the peg placement. For KT, no significant difference was observed between the two groups across all surgeme transitions ($P > .05$).

Table 5 presents the results for the recognition of the trainees' level of experience using the kNN approach. The best results were obtained for k = 5. Although we examined all metrics, only the metrics that showed statistical significance in the group comparison test are included (see **Tables 3** and **4**). For the other metrics the results were much lower so they are omitted. The best performance per evaluation metric is bolded. From **Table 5** it can be seen that the video metrics extracted from the PT task provide better classification. In particular: accuracy 0.86 – 0.81 vs. 0.71 – 0.76, sensitivity 0.81 – 1.00 vs. 0.80 – 0.90, and specificity 0.60 – 0.80 vs. 0.63. Moreover, in both tasks the sensitivity is higher than specificity denoting better performance in the recognition of the residents' class compared to students. Especially the metrics related to the PP surgeme (counts and duration), provide perfect recognition of the residents' class (sensitivity = 1.00). For students, the best performance was yielded by the 'RP duration' metric (specificity 0.80).

Table 3.
Performance Outcomes and Group Comparison for the Peg Transfer Task

	Surgical Residents		Medical Students		*p*-Value
Surgeme (counts)					
Reach for peg	4	[4–5]	7.5	[6–10]	.001*
Transfer peg between graspers	2	[2–2]	3	[2–4]	.075
Place peg into pegboard	4	[4–5]	7	[6–7]	.001*
Surgeme (duration, sec)					
Reach for peg	23.1	[18.3–27.8]	52.8	[45.6–63.7]	.001*
Transfer peg between graspers	16.2	[10.8–19.1]	27.7	[18.3–35.7]	.026*
Place peg into pegboard.	41.7	[34.5–49.6]	90.2	[55.8–98.1]	.001*
Penalty event (counts)					
Peg is dropped	0	[0–1]	3.5	[2–6]	.001*

Values in brackets denote the range between the 25th and the 75th percentile.
*Denotes statistical significance.

DISCUSSION

This study presents a methodology for performance comparison and experience level recognition of surgical trainees, based on video task annotation. The main idea was based on the development of a surgical motion vocabulary of surgical gestures (surgemes) and time-point events. This vocabulary was utilized for annotating the videos of laparoscopic training tasks performed by students and residents. Various metrics were extracted for analyzing task performance. Regarding the comparison of the two groups, our results showed that in the simple PT task, surgical residents performed fewer and shorter surgemes and made fewer errors compared to medical students. In particular, the residents were faster in reaching the peg and placing the peg on the pegboard, denoting greater

Table 4.
Performance Outcomes and Group Comparison for the Knot Tying Task

	Surgical Residents		Medical Students		*p*-Value
Surgeme (counts)					
Reach for needle	1.5	[1–2]	2	[1.2–3]	.159
Orient needle	1	[1–2]	2	[1–3]	.226
Making C-loop around right grasper	1	[1–1]	1	[1–1]	.188
Reach for suture with right grasper	1	[1–1]	1	[1–1]	>.05
Pull suture with both graspers	1	[1–1]	1	[1–1]	>.05
Surgeme (duration, sec)					
Reach for needle	29.9	[24.7–36.2]	50.3	[33.2–73.5]	.022*
Orient needle	9.5	[5.4–13.5]	25.3	[11.2–41.6]	.018*
Making C-loop around right grasper	11.5	[8.2–18.3]	21.4	[11.2–60.7]	.098
Reach for suture with right grasper	5.1	[2.3–6.6]	5.9	[4.6–7.1]	.323
Pull suture with both graspers	3.8	[2.5–9.8]	9.5	[5.4–10.1]	.192
Penalty event (counts)					
Needle is dropped	0.5	[0–1]	1	[0.25–2]	.118

Values in brackets denote the range between the 25th and the 75th percentile.
*Denotes statistical significance.

Figure 5. Statistical comparison results (p values) based on the surgeme transition counts for the peg transfer task, and the corresponding transition diagram. Bold values and arrows with increased weight, denote a significant difference in the transition counts between the two groups.

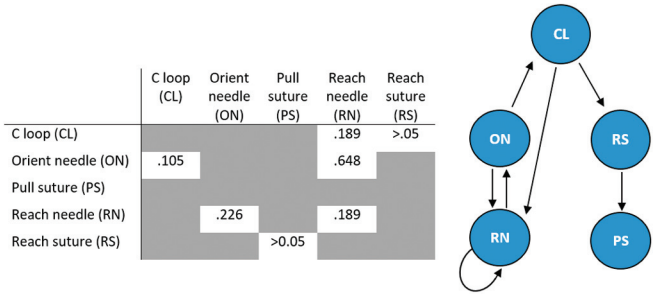

Figure 6. Same as **Figure 5** but for the knot tying task.

dexterity in the performance of these gestures. Moreover, residents were more accurate in the placement/transfer of the pegs (fewer 'drop peg' events). The KT task was more complex and required greater dexterity compared to PT. Among the five surgemes in which the overall task was decomposed to, residents outperformed students only in two surgemes. In particular, residents were faster than students in reaching the needle from the floor of the cavity and orienting the needle. The greater experience of the residents compared to students is reflected in the performance of these two gestures. However, in the other gestures (e.g. 'making a c-loop'), the two groups had similar performance, probably due to the greater dexterity (and thus experience) required to perform these gestures.

In addition to the individual surgemes, we also compared the sequential transitions performed by the two groups. This type of analysis allowed the development of transitions diagrams, which provided a useful tool for assessment of task performance. The transition diagrams showed no performance difference between the groups for the KT task, whereas for PT there was a significant difference in the transitions between the RP and PT surgemes. In particular, the residents performed fewer of these transitions (in both directions) than students, denoting better task performance. The reason that the KT surgeme transitions did not differentiate the two groups may be the number of sequential transitions required to complete this task. For KT the number of sequential surgemes that had to be completed for successful task performance was much less compared to PT. In particular, PT required the performance of 10 sequential surgemes, whereas KT required only five sequential surgemes. Hence, for PT there was a greater chance to find a significantly different number of surgeme transitions between the two groups compared to KT.

Another important finding of this study was that the metrics extracted from the video annotation could be utilized for recognizing the experience level of surgical trainees. Our results showed that for the PT task, using a basic classification algorithm such as kNN, the accuracy in the classification of the two groups was close to 0.86 for one metric (PP duration) and > 0.81 for four metrics. Among the various metrics, PP and RP durations yielded the highest performance: accuracy 0.86 and 0.81, sensitivity 1.00 and 0.81, and specificity 0.70 and 0.80, respectively. The higher sensitivity compared to specificity means that our method was more confident in the classification of the residents. This may be explained by the fact that the metric values of the residents expanded across a limited range compared to students, something that is confirmed by their small interquartile range. Hence, the possibility of misclassifying a student is greater when using a nearest neighbor technique. More advanced machine algorithms may lead to even better results, so this issue certainly deserves further investigation in the future.

Our study also demonstrated that video annotation is a promising tool to evaluate differences in task performance between subjects with variable experience in laparoscopic surgery. Previous studies have employed VR simulation systems and specialized motion tracking devices to evaluate performance differences.[8] The former approach may be tailored to the training requirements, providing plenty of training scenarios and performance metrics, although at a significantly higher cost compared to box trainers. Alternatively, motion tracking devices have been utilized for skills assessment in a box trainer environment. These devices are more cost effective than VR simulators, but they require investment of resources and expertise in hardware installation. Moreover, both approaches provide generic metrics (e.g. tool pathlength) that do not reflect the individual gestures performed during training.

The proposed skills assessment methodology is cost effective as the main set-up is based on a standard box trainer whereas the video annotation software is freely available.

Table 5.
Experience Level Recognition Results Based on k-Nearest Neighbor (k = 5)

	Accuracy	Sensitivity	Specificity
Peg Transfer Task			
Reach for peg (counts)	0.81	0.91	0.70
Place peg on pegboard (counts)	0.81	1.00	0.60
Reach for peg (duration)	0.81	0.81	0.80
Transfer peg between graspers (duration)	0.76	0.91	0.60
Place peg into pegboard (duration)	0.86	1.00	0.70
Peg is dropped (counts)	0.81	0.91	0.70
Knot Tying Task			
Reach for needle	0.76	0.90	0.63
Orient needle	0.71	0.80	0.63

Moreover, no special skills for video annotation are required. In addition, the video annotation process is not affected by the presence of preceptors during the execution of the tasks. The only requirement is the surgical motion vocabulary (surgemes and time-point events), which is defined once for a particular surgical task. The vocabulary may be easily developed by a surgical expert according to the preferred level of gesture granularity. The proposed technique may also be used to guide trainees on how certain surgical gestures, or the task as a whole, could have been performed. For example, given the output from the annotation software, a post-processing technique may inform the trainee about metrics of his/her performance that are below a certain threshold (e.g. median of the corresponding experts' metrics). One limitation of the video annotation methodology is the time required to annotate the videos. However, the employed video annotation software provides plenty of shortcut keys to accelerate this process. Hence, the reviewer may easily perform the annotation online (i.e. while watching the video), without the need to pause and resume the video. Based on our experience, it took about one minute more than the length of the video to complete the annotation. In particular, after familiarization with the shortcut keys of the software, the reviewers performed the annotation almost in parallel with watching the video of the task.

Although this study aimed to present proof-of-concept experiments for surgical performance recognition based on video-assisted annotation of surgical tasks, a basic limitation is the absence of an expert group. Preliminary studies on a limited sample (five experts) have shown that experts performed the tasks (especially KT) with significantly shorter surgemes and fewer surgeme transitions compared to the other groups. However, a larger sample is required to derive definite conclusions, along with performance data from other video analysis methods. Currently we are in the process of developing a fully annotated dataset (trainees with variable experience, surgeme annotation, OSATS evaluation, etc.) that includes synchronized video and kinematic data from a range of surgical tasks performed on a box trainer. Our aim is to offer this dataset available to the scientific community for further comparative evaluation among various measures/techniques of surgical performance analysis (e.g. checklists, video and gesture analysis, combination of them).

CONCLUSION

In this article we propose a methodology for skills assessment and performance analysis of surgical trainees based on video annotation of laparoscopic tasks. The proposed method provides a tool not only for performance comparison among different experience groups, but also for extracting interpretable metrics of surgical performance. Our findings also showed that these metrics are promising in the recognition of the experience level of surgical trainees. In the future we aim to expand the surgical motion vocabulary and apply the technique for assessment of laparoscopic tasks performed in the operating room.

References:

1. Darzi A, Smith S, Taffinder N. Assessing operative skill. Needs to become more objective. *BMJ.* 1999;318(7188):887–888.

2. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg.* 1997;173(3):226–230.

3. Bhatti NI, Cummings CW. Competency in surgical residency training: defining and raising the bar. *Acad Med.* 2007;82 (6):569–573.

4. Martin J, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg.* 1997;84(2):273–278.

5. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg.* 2005;190(1):107–113.

6. Shah J, Darzi A. Surgical skills assessment: an ongoing debate. *BJU Int.* 2001;88(7):655–660.

7. Pellegrini CA. The ACGME outcomes project. *Surgery*. 2002; 131(2):214–215.

8. Reiley CE, Lin HC, Yuh DD, Hager GD. Review of methods for objective surgical skill evaluation. *Surg Endosc*. 2011;25 (2):356–366.

9. Leong JJH, Nicolaou M, Atallah L, Mylonas GP, Darzi AW, Yang G-Z. HMM assessment of quality of movement trajectory in laparoscopic surgery. *Comput Aided Surg*. 2007;12(6):335–346.

10. Loukas C, Nikiteas N, Kanakis M, Georgiou E. The contribution of simulation training in enhancing key components of laparoscopic competence. *Am Surg*. 2011;77(6):708–715.

11. Rosen J, Brown JD, Chang L, Sinanan MN, Hannaford B. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. *IEEE Trans Biomed Eng*. 2006;53(3):399–413.

12. Winkler-Schwartz A, Bissonnette V, Mirchi N, et al. Artificial intelligence in medical education: best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ*. 2019;76(6):1681–1690.

13. Loukas C. Video content analysis of surgical procedures. *Surg Endosc*. 2018;32(2):553–568.

14. Islam G, Kahol K, Li B, Smith M, Patel VL. Affordable, web-based surgical skill training and evaluation tool. *J Biomed Inform*. 2016;59:102–114.

15. Loukas C, Georgiou E. Performance comparison of various feature detector-descriptors and temporal models for video-based assessment of laparoscopic skills. *Int J Med Robot Comput Assist Surg*. 2016;12(3):387–398.

16. Funke I, Mees ST, Weitz J, Speidel S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int J Comput Assist Radiol Surg*. 2019;14(7):1217–1225.

17. Loukas C, Nikiteas N, Schizas D, Georgiou E. Shot boundary detection in endoscopic surgery videos using a variational Bayesian framework. *Int J Comput Assist Radiol Surg*. 2016;11 (11):1937–1949.

18. Zappella L, Béjar B, Hager G, Vidal R. Surgical gesture classification from video and kinematic data. *Med Image Anal*. 2013; 17(7):732–745.

19. Loukas C, Georgiou E. Surgical workflow analysis with Gaussian mixture multivariate autoregressive (GMMAR) models: a simulation study. *Comput Aided Surg*. 2013;18(3–4): 47–62.

20. Zia A, Sharma Y, Bettadapura V, et al. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *Int J Comput Assist Radiol Surg*. 2016;11 (9):1623–1636.

21. Zia A, Sharma Y, Bettadapura V, Sarin EL, Essa I. Video and accelerometer-based motion analysis for automated surgical skills assessment. *Int J Comput Assist Radiol Surg*. 2018;13 (3):443–455.

22. Kipp M. ANVIL-a generic annotation tool for multimodal dialogue. In: *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)* 2001:1367–1370. https://www.anvil-software.org/.

23. Ahmidi N, Tao L, Sefati S, et al. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Trans Biomed Eng*. 2017;64(9):2025–2041.