Modeling Surgical Technical Skill Using Expert Assessment for Automated Computer Rating

David P. Azari, MS, Lane L. Frasier, MD,† Sudha R. Pavuluri Quamme, MD, MS,† Caprice C. Greenberg, MD, MPH,*† Carla M. Pugh, MD, PhD,*† Jacob A. Greenberg, MD, EdM,† and Robert G. Radwin, PhD*‡*

Objective: Computer vision was used to predict expert performance ratings from surgeon hand motions for tying and suturing tasks.

Summary Background Data: Existing methods, including the objective structured assessment of technical skills (OSATS), have proven reliable, but do not readily discriminate at the task level. Computer vision may be used for evaluating distinct task performance throughout an operation.

Methods: Open surgeries was videoed and surgeon hands were tracked without using sensors or markers. An expert panel of 3 attending surgeons rated tying and suturing video clips on continuous scales from 0 to 10 along 3 task measures adapted from the broader OSATS: motion economy, fluidity of motion, and tissue handling. Empirical models were developed to predict the expert consensus ratings based on the hand kinematic data records.

Results: The predicted versus panel ratings for suturing had slopes from 0.73 to 1, and intercepts from 0.36 to 1.54 (Average $R^2 = 0.81$). Predicted versus panel ratings for tying had slopes from 0.39 to 0.88, and intercepts from 0.79 to 4.36 (Average $R^2 = 0.57$). The mean square error among predicted and expert ratings was consistently less than the mean squared difference among individual expert ratings and the eventual consensus ratings.

Conclusions: The computer algorithm consistently predicted the panel ratings of individual tasks, and were more objective and reliable than individual assessment by surgical experts.

Keywords: computer vision, marker-less video tracking, objective structured assessment of technical skills, objective structured assessment of technical skills, surgical task analysis

(*Ann Surg* 2019;269:574–581)

Computer vision technologies utilize digital cameras and computer algorithms for tracking motions, classifying information, detecting features, and recognizing patterns in digital images and videos.[1] It has impacted a diverse field of applications, ranging from industrial robotics, intelligent and autonomous vehicles, security surveillance, manufacturing inspection, and human-computer interaction. Applications include face detection, expression and emotion recognition, video surveillance and monitoring, and motion tracking and analysis,[2,3] to name a few. This proof-of-concept article compares models using computer motion tracking of surgeon hand movements against performance ratings made by experts.

A growing body of literature suggests that the technical skill of the surgeon can affect patient outcomes.[4–6] Technical skill contributes to three-quarters of adverse events in the operating room,[7,8] but common methods of measuring performance (ie, in-training reports and mentor observations and evaluation) have been criticized for lacking reliability and being generally subjective[9] and are under increasing pressure to document proficiency.[10] The objective structured assessment of technical skills (OSATS) is based on observing performance in real-time and rating candidates along a series of Likert-based scales in conjunction with a procedure-specific checklist.[11] Studies examining OSATS have demonstrated strong validity evidence in accordance with Kane's framework,[12,13] especially in providing formative feedback during training.[14] Despite this proven track record, correctly implementing OSATS is resource-intensive and time-consuming,[15] prompting exploration of more efficient assessment techniques.[16,17]

Motion capture and tracking of surgeon hand movements have the potential to fill these gaps.[18–22] The Imperial College Surgical Assessment Device (ICSAD) is an excellent example. Studies utilizing ICSAD have shown it is possible to identify consistent differences in trainees and experts by observing the position of small sensors placed on the hands,[23,24] and that these movements correlate strongly with OSATS global rating assessments.[25] Unfortunately, this self-contained system is limited to benchtop models and consequently not easily applied in the operating room.[26] Assessment in the operating room requires a noninvasive and scalable means of observing surgical motion without relying on embedded sensor technology.

Video motion capture of the surgeon's hands in the surgical field, facilitated by increasing availability of cameras in the OR, offers an expedient alternative to integrated sensor-software systems. We have developed marker-less video processing methods using conventional single camera digital video to reliably track the motion trajectory of a selected region of interest over successive video frames without the need for sensors or markers.[27–29] We have also demonstrated how tracked hand motion can quantify kinematic properties of movements and exertions during specific tasks.[27,30,31] Previous studies by our group have used this technology to isolate kinematic differences in surgical hand motion,[20] and identify statistically significant differences between surgeon roles (attending vs resident), tasks (tying vs suturing), and tissue types during open surgery.[32]

This study compares expert ratings of surgical skills (our current gold standard) to kinematic measures of surgeon hand motions to evaluate the potential use of video to automatically measure technical skill. The objective is to establish empirical models of expert-rated in vivo performance during operations by

From the *Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI; †Wisconsin Surgical Outcomes Research (WiSOR) Program, Department of Surgery, University of Wisconsin-Madison, Madison, WI; and ‡Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI.

The project described was supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. Lane Frasier is currently supported by AHRQ F32 HS022403 and also received support from NIH T32 CA90217 and the AAS Research Fellowship Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

The authors report no conflicts of interest.

Reprints: Robert G. Radwin, PhD, University of Wisconsin-Madison 1550 Engineering Drive Madison, WI 53706. E-mail: rradwin@wisc.edu.

ISSN: 0003-4932/17/26903-0574
DOI: 10.1097/SLA.0000000000002478

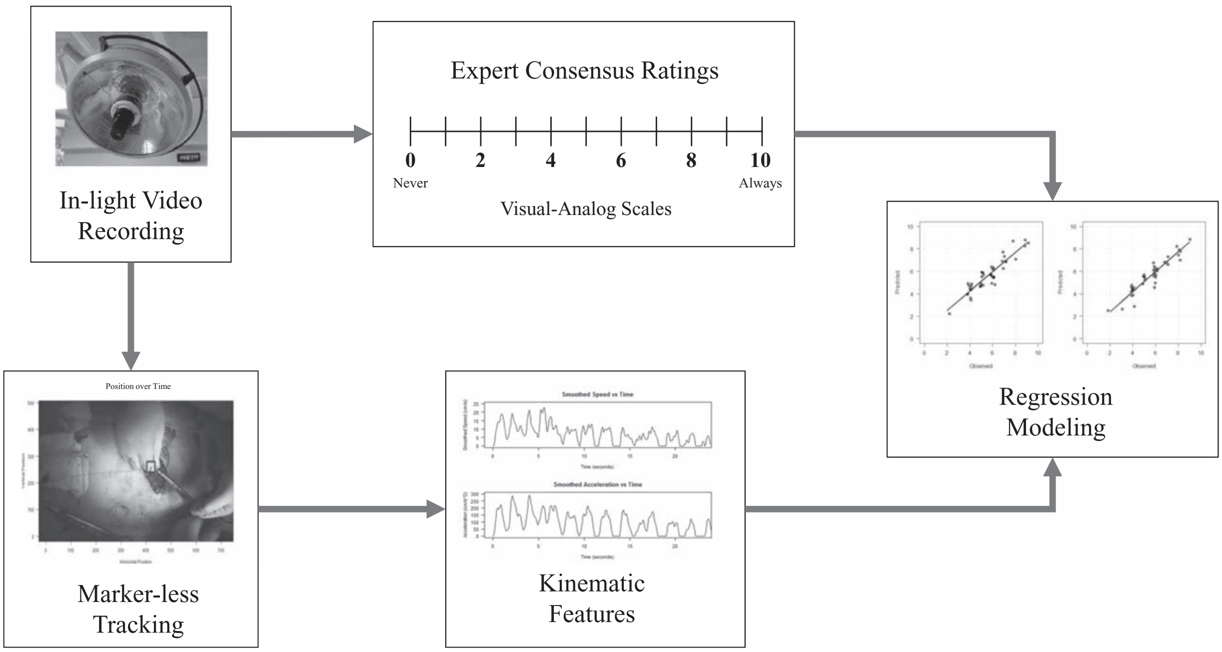

FIGURE 1. Study overview. Video from the in-light camera in the OR was collected. A panel of 3 expert surgeons rated task performance through a consensus process. Hand motion in the video was tracked and a series of kinematic feature characteristics were extracted. Regression models were used to predict the expert consensus ratings based on the extracted features for all the video clip segments.

extracting kinematic features of tracked hand movements. We created task-specific rating scales and tracked hand motion records to predict subject matter expert ratings of surgical performance for a series of suturing and tying tasks. We hypothesize that the kinematic features of a surgeon's hand motion measured using marker-less tracking can be used to accurately model subjective performance ratings made by a panel of experts.

METHODS

Participants

This study utilized in-light OR video cameras (Fig. 1) to capture hand motions of 9 surgeons (6 attendings and 3 residents) during 16 surgical cases. Cases included colorectal, complex upper gastrointestinal, hepatobiliary, surgical oncology, transplant, vascular, thoracic, and cardiac operations. Our institutional review board granted ethical approval for recording and analyzing video data for these operations. Written informed consent was obtained from participant surgeons in advance. Cases were recorded during 9 months. Owing to the positioning and limited field of view of the in-light OR camera, no patient details or protected health information were captured. Audio was not recorded.

Video Selection

Recording of approved cases was initiated remotely after the operation began and stored on a secure hospital computer enabled with a video encoder (AXIS video encoder, Axis Communications, Lund, Sweden). We reviewed the videos using Multimedia Video Task Analysis (MVTA) software (Wisconsin Alumni Research Foundation, Madison, WI) developed by Yen and Radwin.[33] MVTA allows real-time review, labeling, and exporting of video events. A member of the research team familiar with operative technical tasks (LLF) screened and categorized the videos in MVTA for segments of tying or suturing tasks wherein the hands were clearly visible for at least 5 seconds. Suturing tasks were categorized as bowel anastomosis (SBA), complex anastomosis (SCA) such as hepaticojejunostomy, or suturing on the body wall (SBW) during closure or stoma formation, for example. Each tying task was categorized as follows: closing and external tying on the body wall (TBW), intra-abdominal superficial tying (TSF), or intra-abdominal deep tying (TDP). A full taxonomy and further information on the surgical procedures performed are described by Frasier et al.[32]

Rating Scales

Subjective visual-analog rating scales[34] were created for evaluating surgical performance during short clips (5–30 seconds) for motion economy, fluidity of motion, and tissue handling (Fig. 2). We developed the scales utilizing the OSATS motion scales (ie, respect for tissue, time and motion, and instrument handling) as assessment blueprints. The scales were framed to evaluate performance during a short clip rather than for an individual during the whole procedure (thereby omitting procedure-specific checklists), and defined to comprise the entire range (0–10) of possible behaviors observed during that segment.

Fluidity of motion is a measure of hesitancy, pauses, or changes in direction and "resets," which may be a component of Moulton's "slowing down,"[35] or contribute to time spent idle.[36] Tissue handling quantifies the appropriateness of the surgeon's force and tension when manipulating the tissue,[37,38] and varies based on the tissue's friability and fragility.[18] Motion economy is defined as efficiency of movement, or conservation of energy in any trajectory. Such behavior is consistently documented as a mark of expert psychomotor behavior, and increasingly studied as a measure of surgical skill.[25,39]

A consensus panel of 3 expert surgeons (CCG, JAG, CMP) viewed each clip in random order and independently rated hand motion across the scales. Each surgeon announced their ratings to the

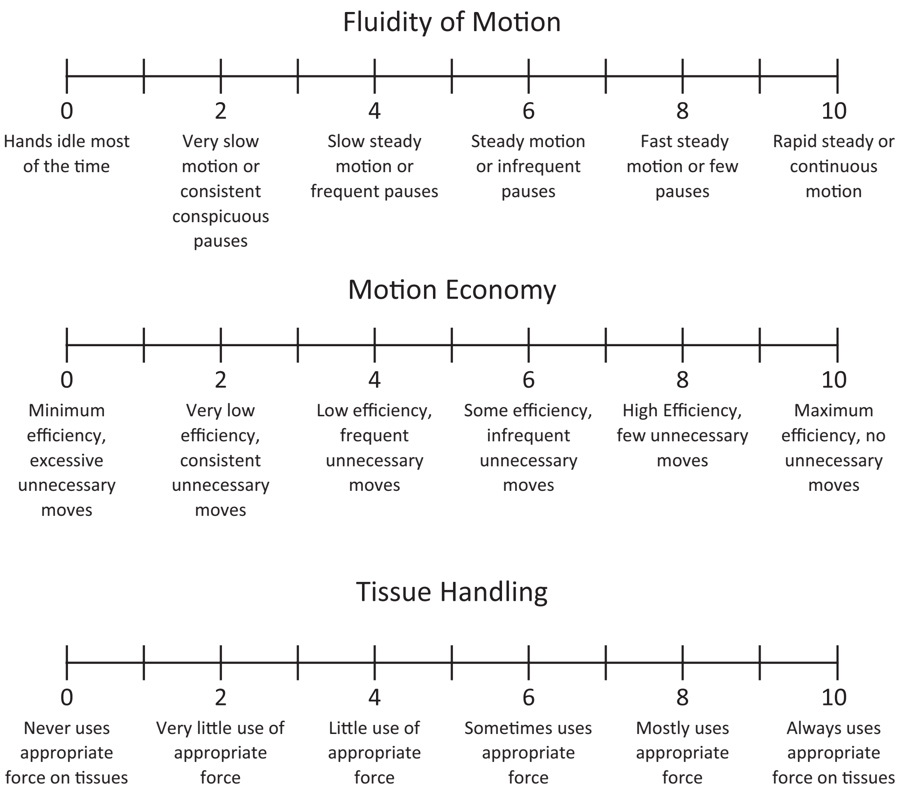

FIGURE 2. Subjective rating scales were based on fluidity of motion, motion economy, and tissue handling for tying and suturing tasks.

group. Discrepancies were discussed until consensus was achieved for each rating. If absolute consensus was unable to be achieved, the clip was scored according to the majority evaluation.

Motion Tracking

Custom video tracking software,[27] developed in one of the authors' (RGR) lab, was used to trace a region of interest (ROI) on a visible portion of the surgeon's hands (generally the index finger or thumb) across successive video frames. We previously used this marker-less tracking approach to examine differences in hand motion between attending and resident surgeons,[32] to discriminate between dominant and nondominant hand motion during reduction mammoplasty[20] and to evaluate hand-motion patterns during simulated clinical breast exams.[31] The *x-y* pixel location of the ROI for each frame (every $\frac{1}{30}$ of a second) was recorded. An analyst selected the size and position of the ROI to track the surgeon's hand for each video clip, and because of occasional occlusions of the hands and changes in lighting, supervised tracking of the ROI, making manual corrections as necessary.

Calibration

In-frame visible measurements of the hands were used to calibrate each video clip from pixels to millimeters. We previously used hand dimensions to provide acceptable estimates of hand speed.[30] Observed proximal interphalangeal joint breadth was scaled to the population means of males (23.0 mm) or females (19.9 mm) depending on the sex of the surgeon. The proximal interphalangeal joint breadth was selected because of its small coefficient of variation for males (0.071) and females (0.064) as determined by the US Army.[40] The video was calibrated based on average hand measurements across 3 different frames, and recalibrated for any change in the position of the in-light camera or the surgical field.

Variable Selection

The tracked record of the ROI location allowed quantification of instantaneous displacement, speed and acceleration of the surgeon's hand, and several additional measures including jerk[41] and spatiotemporal curvature.[42] Jerk is the third derivative of position with respect to time and generally represents how smooth a motion is, whereas the spatiotemporal curvature function is a measure of direction change based on multiple derivatives of the position signal and is used to indicate the number of discrete movements.

A second-order Butterworth filter within empirically observed upper limits of hand frequency for mono-hand tasks[43] (pass band = 0.005–1.000 Hz) was applied to smooth the acceleration and curvature signals and a Fast Fourier Transform (FFT) function was applied to each signal set to isolate any consistently cyclic and repeated motion patterns, a growing avenue of research in surgical dexterity.[44] Moving averages were calculated for the original and smoothed signal types (speed, acceleration, curvature), as well as relative densities to see how much area is consistently and repeatedly traversed by the hand (this can essentially be interpreted as a proxy for path length, a common variable output by platforms such as the ICSAD[24]). These predictor variables are distributed across the variable families shown in Table 1.

Modeling Process

We developed a set of linear regression models to test whether the kinematic features could predict the expert ratings across each of the motion scales. The predictor variables were examined for collinear relationships, and subsets of these selected for regression analysis. In selecting variables to initiate the modeling process, we hypothesized that the average speed and consistent locations in the speed signal would give strong predictions of motion economy

TABLE 1. Features Used for Modeling the Consensus Rating Scales

Summary kinematics	Mean, median, and max of speed, acceleration, curvature of motion, and jerk
RMS speed and acceleration	Root mean-squared transformation of speed; acceleration (respectively)
Moving averages	Simple, exponential, and moving averages applied to original and smoothed speed, acceleration and curvature of motion signals
Peak counts	Number of peaks, applied to pure and filtered speed, acceleration, curvature of motion, and all moving averages at thresholds at intervals of 20% from 0% to 100% of maximum signal amplitude and intervals of 20 mm/s for speed (up to 100) and 200 mm/s^2 for acceleration (up to 1000)
Peak frequencies	Number of peaks divided by length of video clip (s) for each type of peak count category
Peak variation	Coefficient of variation in peak arrivals (standard deviation/mean) for each peak count category
Idle time (%)	Percent of time spent below speed threshold (set every 5 mm/s up to 50 mm/s)
Working area	Mean variance in *x-y* distance calibrated positions over all recorded frames
Path density (avg, med, SD, max)	Ratio of recurrent (range 0–7) recorded *x-y* positions to total recorded *x-y* positions
Speed density (avg, med, SD, max)	Ratio of instantaneous speed in recurrent (range 0–7) recorded *x-y* positions to total recorded *x-y* positions
Curvature density (avg, med, SD, max)	Ratio of instantaneous curvature of motion values in recurrent (range 0–7) recorded *x-y* positions to total recorded *x-y* positions

Kinematic variable family names (left) and associated descriptions (right). RMS indicates root mean square; SD, standard deviation.

ratings, whereas the peak arrival rate in the acceleration signal would predict fluidity ratings and jerk or pauses would predict tissue handling ratings. In other words, motion economy might correspond to how fast the surgeons are moving in a consistent area, whereas fluidity might relate to how many changes in speed there are, and tissue handling would be sensitive to any sudden, repeated changes in direction. The models were optimized using the Akaike information criterion to balance the prediction against the number of variables.[45] Every consensus rating scale (fluidity of motion, motion economy, tissue handling) was first modeled for each distinct category of surgery (ie, SBW, SBA, SCA, TBW, TSF, TDP). We then examined the combined predictions from suturing tasks and tying tasks for each rating scale (Fig. 3).

Validation

To assess the internal validity of the prediction models, we compared the sum of squared errors (SSE), a statistical measure of variance, to the leave-one-out predicted residual sum of squares (PRESS) statistic—a common approach in validating models where test data is not available.[46,47] The SSE measure assessed the overall fit, whereas the PRESS statistic penalizes the model if it depends on outliers. The closer these values are, the better the model matches the current data without relying on outliers, and the better prediction it provides (Table 2). We calculated the root-mean SSE and root-mean PRESS statistics to normalize the comparison between models, and refer to this value as the model's ''error'' for ease of reference. Ideal

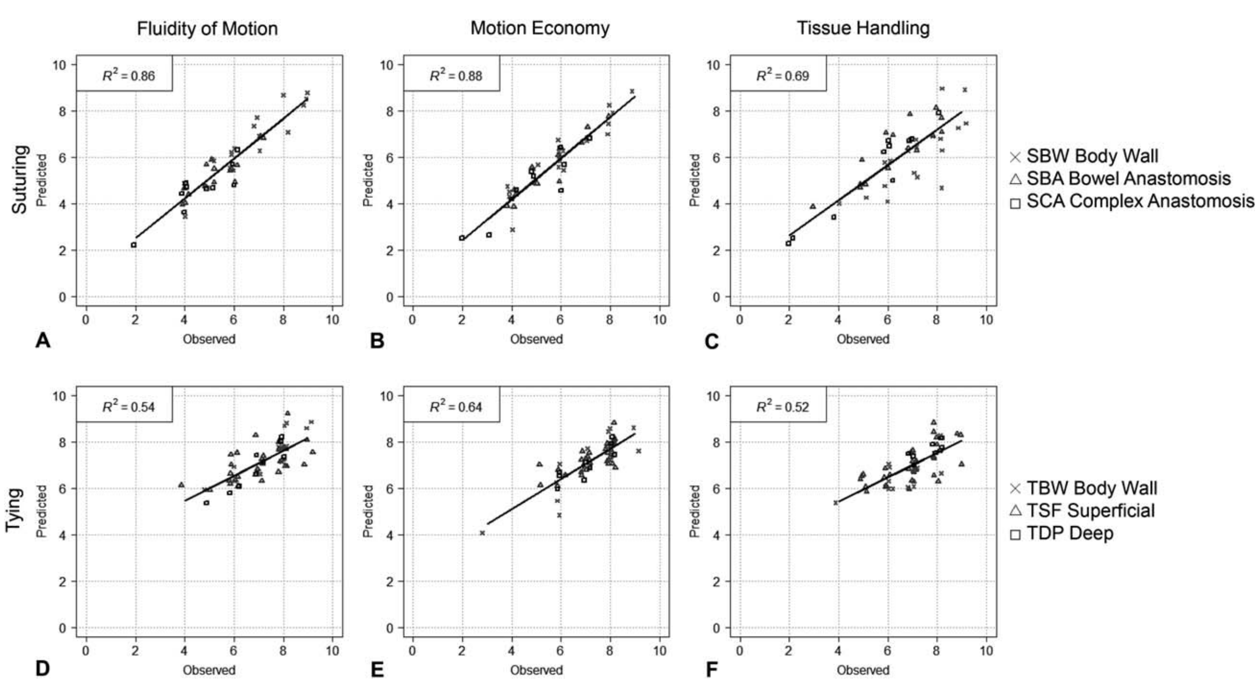

FIGURE 3. Prediction models vs observed results for suturing and tying tasks across all tissue types and scales for (A) suturing fluidity of motion, (B) suturing motion economy, (C) suturing tissue handling, (D) tying fluidity of motion, (E) tying motion economy, and (F) tying tissue handling.

TABLE 2. Regression Model Summary Statistics and Predictor Variables

	Fluidity of Motion			Motion Economy			Tissue Handling		
Task	Pred vs Obs	Variables	P	Pred vs Obs	Variables	P	Pred vs Obs	Variables	P
Suturing SBW (n ¼ 19)	m U 0.85 b U 0.99 R^2 U 0.82	Peak acceleration (T U 5000) Peak smoothed speed (T U 100) Peak curvature (T U 0)	0.000 0.025 0.006	m U 0.87 b U 0.78 R^2 U 0.85	Peak curvature (T U 0) Peak acceleration (T U 6000) Peak speed (TU 100)	0.001 0.000 0.005	m U 0.82 b U 0.30 R^2 U 0.84	Peak variance of acceleration (T U 3000) Smoothed peak curvature (T U 0.5) RMS acceleration Peak speed (TU 100)	0.008 0.000 0.000 0.000
Suturing SBA (n U 15)	m U 0.73 b U 1.43 R^2 U 0.62	Peak acceleration (T U 1000) RMS speed Peak rate speed (T U 1000) Median speed	0.021 0.008 0.027 0.149	m U 0.89 b U 0.59 R^2 U 0.85	Path density of median speed (R U 7) Peak acc. (TU 9000) Peak acc. (TU 1000) FFT of Speed	0.061 0.013 0.005 0.106	m U 0.75 b U 1.54 R^2 U 0.66	Path density of mean speed (V U 7) Peak acceleration (T U 9000) Median acceleration Peak curvature (TU 0)	0.036 0.017 0.118 0.191
Suturing SCA (n U 10)	m U 0.75 b U 1.13 R^2 U 0.63	Weighted avg. of smooth speed (W U 18s) Path density of average curvature (T U 0) Peak variance of speed (T U 300)	 0.061 0.074 0.582	m U 0.85 b U 0.75 R^2 U 0.77	Path density of max. curvature (T U 0) Peak variance of speed (T U 300) Peak acceleration (T U 8000)	0.184 0.050 0.020	m U 0.93 b U 0.36 R^2 U 0.90	Weighted average of smoothed curvature [0:1] (W U 53s) Peak rate of speed (T U 100) Path density of standard deviation in curvature (R U 2)	0.007 0.001 0.050
Tying TBW (n U 15)	m U 0.73 b U 2.05 R^2 U 0.65	Weighted average of acceleration Fast Fourier transform (FFT) of speed RMS speed	0.001 0.004 0.000	m U 0.76 b U 1.65 R^2 U 0.70	Weighted average of curvature Peaks of curvature [0:1], (T U 80% of Maximum) Idle Time (T U 50%)	0.000 0.011 0.005	m U 0.57 b U 2.88 R^2 U 0.46	Peak variance of speed (T U 700) Maximum speed Peak acceleration (T U 1000)	0.004 0.040 0.092
Tying TSF (n U 35)	m U 0.39 b U 4.36 R^2 U 0.33	Peak acceleration (T U 1000) Peak speed (TU 500) Idle time (T U 30%)	0.073 0.727 0.047	m U 0.46 b U 3.96 R^2 U 0.40	Peak acceleration (T U 9000) Maximum Jerk Working area	0.009 0.005 0.015	m U 0.45 b U 3.83 R^2 U 0.40	Weighted average of smoothed curvature Path density of position (R U 7) FFT of acceleration	0.0005 0.017 0.040
Tying TDP (n U 9)	m U 0.88 b U 0.79 R^2 U 0.85	Peak speed (TU 600) Peak speed (TU 300)	0.028 0.199	m U 0.74 b U 1.81 R^2 U 0.65	FFT of speed Working area	0.018 0.334	m U 0.56 b U 3.37 R^2 U 0.30	Peak variance of speed (T U 0) Path density of med. speed (R U 7) FFT of acceleration	0.132 0.834 0.131

Acceptable models (slope between 0.5 and 1.5, with an intercept of 2 or less and R^2 0.75) are indicated in bold text.

Pred U predicted, Obs U observed, m U slope, b U intercept, with predicted U m (observed) R b; R^2 U adjusted coefficient of determination.

Avg. indicates average; Max., maximum; Med., median; R, number of recurrent position traversal as threshold for density calculation; RMS, root mean square; T, threshold in mm/s for speed or mm/s^2 for acceleration; W, window.

prediction models would have several properties. The predicted and ground truth ratings should form a straight line from (0,0) to (10,10), such that any unitary increase in expert appraisal would be met with a similar predicted increase. We arbitrarily defined acceptable models to have a slope between 0.5 and 1.5, an intercept within 2 of zero, and an $R^2 > 0.75$. Although other thresholds could be defined, these bounds were chosen to ensure a limited difference between the ground truth and the prediction, and a prediction consistent throughout the domain of each scale.

RESULTS

Video Data

In total, 103 video clips (mean time = 11.72 seconds) were recorded of 6 attending and 3 resident surgeons performing suturing and tying tasks throughout 16 varied operations. These included SBA, SCA, and SBW suturing clips (n = 44), and TBW, TSF, and TDP tying clips (n = 59). The marker-less hand tracking frame-by-frame position data provided >1500 kinematic variables, spread across the variable types in Table 1.

Task Expert Rating Scales

Each clip was observed by 3 expert surgeons and rated along 0 to 10 analog scales for motion economy, fluidity of motion, and tissue handling. The expert raters achieved consensus after a maximum of 3 iterations for all the video clips. Observed ratings for suturing tasks ranged from 2 to 9 (mean = 5.91, SD = 1.62). Tying task ratings ranged from 3 to 9 (mean = 7.12, SD = 1.10) and were particularly skewed toward the upper end of the scale.

Prediction Models of Expert Ratings

The linear regression slopes and intercepts for each of the rating scales and task types are seen in Table 2. The models for fluidity of motion for suturing (Fig. 3A; slope = 0.86, intercept = 0.80, R^2 = 0.86) and motion economy for suturing (Fig. 3B; slope = 0.89, intercept = 0.64, R^2 = 0.88) had the best predictions, and moderately better than tissue handling of suturing (Fig. 3C; slope = 0.76, intercept = 1.1, R^2 = 0.69).

The predicted versus consensus ratings of motion economy for tying tasks (Fig. 3E; slope = 0.65, intercept = 2.51, R^2 = 0.64) had better predictions than tissue handling (Fig. 3F; slope = 0.53, intercept = 3.33, R^2 = 0.52) and fluidity of motion (Fig. 3D; slope = 0.54, intercept = 3.28, R^2 = 0.54).

Prediction Model Validity

Linear prediction models were validated by comparing their difference, or "error," between root-mean predicted residual (PRESS) and root-mean sum of square errors (SSE). Motion economy had the least error (0.17, 0.07) during bowel anastomosis and superficial tying, respectively, as well as the least average error (0.27) for suturing ratings overall. Fluidity of motion models similarly performed well for suturing and tying along the body wall with errors of 0.21 and 0.11. Tissue handling had the highest error for both suturing and tying tasks (0.52, 0.26). Low errors suggest strong positive relationships between hand kinematics and expert ratings—a necessary component of Kane and Messick's modern approach to validity.[13]

DISCUSSION

This study demonstrates that computer vision capturing kinematic data from marker-less motion tracking of video records offers an objective and scalable approach for measuring surgical skill, and establishes evidence consistent with Kane's validity framework.[13] We created subjective rating scales for fluidity of motion, motion economy, and tissue handling using existing OSATS measures as an assessment blueprint. We used kinematic features of the hands to develop prediction models of the expert ratings and compared the model performance against the variance among experts. Models consistently had less variance than the individual experts exhibited before consensus (Fig. 4). In this process, we identified the kinematic measures most closely linked with expert surgeons' ratings of common surgical tasks. Because this approach relies only on noninvasive video tracking and not invasive markers for motion capture, it

FIGURE 4. Comparison of variance in video-based model (mean square error) and individual prediction ratings (mean squared difference) from the expert consensus ratings. The model consistently provides greater reliability than the individual experts in predicting the consensus ratings.

represents a critical step forward in a scalable and reproducible method for objective assessment of surgical technical skill during actual, open operations.

Numerous features were extracted from the video. Hand position, speed, acceleration, and curvature of motion were measured and recorded. Based on previous expectations of psychomotor performance, a series of additional variables were calculated, including peak frequency and variance, path density, working area, and moving averages. Peak arrival rates in both the unfiltered and smoothed speed and acceleration signals were consistently correlated and significant in the prediction models.

The fluidity of motion and motion economy models for all suturing tasks had slopes between 0.73 and 1, and intercepts between 0.30 and 1.54. Although models of tissue handling underperformed the former, the tissue handling prediction for complex or bowel anastomosis suturing tasks had a slope of 0.93 and intercept of 0.36. Many models utilized the peak rate arrivals in the acceleration signal, and achieved slopes close to 1, with intercepts between 0.5 and 1.5, which indicates excellent fit. These results support the conclusion that kinematic features of a surgeon's hand motion, measured using marker-less tracking, can accurately model subjective performance ratings made by a panel of experts. Such measures are useful in providing objective and automatic feedback, and necessary in developing evidence of validity for future competency-based assessments.[48]

There are several limitations to consider. First, this study was contingent on access to video of live operations. Thus, not all scores were observed across all task subtypes. Tying tasks demonstrated skewed expert ratings or had a limited range of scores, reducing the variance and making these tasks more difficult to predict, despite low error estimates. Additionally, the task-specific rating scales themselves did not always address the surgical context. For example, a surgeon with excellent fluidity may exhibit periods of frequent pauses to gather information, avoid distraction, compensate for a hand tremor, or simply resolve confusion or nervousness. Although the detection itself does not suggest a particular decision or course of action to improve, automatically identifying events such as "slowing down," as Moulton[35] describes, could streamline video analysis to target critical points in an operation and provide insight into surgical decision making. In other words, computer vision technology might automatically detect difficulties in an operation, demonstrated by acute changes in hand motion.

This technology also ignores the consequences to a procedure or a patient, and would need to meet additional evidence requirements as outlined by Kane and Messick[12,13] before any deployable assessment routine takes shape. Such assessment would also need to incorporate whether the surgeon identifies the anatomy and adjusts their technique appropriately, commits errors, or generally performs the operation correctly. A suturing task with low fluidity of motion (high hesitancy, frequent pauses) may exhibit significant motion while repositioning the driver multiple times, without contacting tissue, despite a continuous motion tracking record. It is possible that such events may be isolated by combining the different scale predictions (ie, sudden poor tissue handling during a period of low fluidity and high motion economy may hold valuable information about the current state of the surgery.) Future work may address these challenges by considering the implications or consequences of assessment scores for a whole surgical case. These investigations may consider the quality of observed ties or stitches,[49] and focus on more advanced relationships between kinematic features and the overall surgical state, utilizing patterns of motion across repeated cycles[21] and language models[26,50] to automatically classify surgeons and procedures.

We anticipated that fluidity would produce the best prediction outcomes based on recorded speed, and that tissue handling would be most difficult to model, given dependence on tissue type. Similarly, the more complex anastomoses and tying tasks were expected to have greater variability in hand movements—potentially making them more difficult to predict. The tissue handling model for anastomoses (bowel and complex) performed better than tissue handling along the body wall, supporting the hypothesis that tissue handling ratings may be more sensitive to the categorization process and exhibit higher variation. It is also possible that the cues the raters used to gauge tissue handling were not as readily identifiable in the motion tracking record for suturing along the body wall (ie, hand shape, individual finger dexterity), and therefore currently unavailable to the automatic routine.

Arranging time for the expert panel to view and rate the videos was the most difficult aspect of the study. Surgeons reported achieving consensus relatively easily after discussion, but had general difficulty when the field of view was limited. How surgeons resolved disagreements for these clips may suggest additional contextual features to improve the predictions, and may help to establish evidence of validity in the response process itself. Collecting and applying motion-tracking routines to the videos played in real-time was also resource-intensive and time-consuming. More complex procedures with multiple changes in patient or camera positions, shifting light conditions, and occlusions required active input from human observers to recalibrate and instantiate a new tracking ROI. Future software could expedite this workflow, allowing observers to adjust tracking properties on the fly. Software may also allow surgeons to rate videos independently or remotely and resolve any discrepancies in a consensus rating over time. As video collection and editing programs advance, and video recording of operations becomes more widespread, we anticipate that these processing stages will require less manual effort.

In future studies, the overall difficulty with skewed distributions could be mitigated by collecting videos of specific clinically simulated scenarios, wherein variety and experience are intentionally controlled to accommodate the full range of performance ratings. In this setting, the rating scales and kinematics could incorporate different surgical approaches, as some surgeons throw 1-handed ties, whereas others prefer 2-handed ties and the kinematic path for these are quite different, limiting the features available for comparison. An expert viewer may be able to distinguish between a successfully thrown knot and a dropped ligature, but depending on how the surgeon moves after the error, their kinematic path may look identical. Such situations continue to require subjective supervisory intervention and assessment, and future work is needed to automatically identify these contextual factors, and assess measures of reproducibility among different stations and raters in more controlled (ie, benchtop) simulations.

Access to immediate, reproducible kinematic-based feedback based on video review can inform self-assessment, direct practice of specific tasks, and build overall surgical skill. This may help to provide a venue in which skill development is quantifiably traceable to training interventions. If used to track expert surgeons, the capacity to deconstruct surgical skill can provide a deeper understanding of the kinematics of surgical performance and aid in the development of novel approaches to skill acquisition. Understanding the components of performance and providing such feedback can potentially shorten the current learning curve and help detect skill decay using objective measures either during periods of inactivity or toward the end of surgical careers.

The findings in this study represent a measurable step forward in creating more objective, reproducible, and accessible assessments of surgical skill commensurate with direct video observation by panel of expert raters. The prediction models have the potential to be packaged into automatic, on-demand feedback in and out of training settings, providing a reliable measure of assessment and consistent feedback to facilitate direct practice of hand motions for specific tasks. Future work is needed to explore the capacity of motion capture alone to make decisions about trainee competence. As video capture is becoming increasingly common in the operating room, computer vision motion tracking allows for a uniquely scalable approach to surgical skill analysis.

REFERENCES

1. Gavrila D. The visual analysis of human movement: a survey. *Comput Vis Image Underst*. 1999;73:82–98.
2. Poppe R. Vision-based human motion analysis: an overview. *Comput Vis Image Underst*. 2007;108:4–18.
3. Wang L, Hu W, Tan T. Recent developments in human motion analysis. *Pattern Recognit*. 2003;36:585–601.
4. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med [Internet]*. 2013;369:1434–1442.
5. Reznick RK, MacRae HM. Teaching surgical skills—changes in the wind. *N Engl J Med*. 2006;355:2664–2669.
6. Darzi A, Smith S, Taffinder N. Assessing operative skill needs to become more objective. *Br Med J [Internet]*. 1999;318:887–888.
7. Rogers SO Jr, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. *Surgery*. 2006;140:25–33.
8. Greenberg CC. Learning from adverse events and near misses. *J Gastrointest Surg*. 2009;13:3–5.
9. Moorthy K, Munz Y, Sarker SK, et al. Objective assessment of technical skills in surgery. *Br Med J [Internet]*. 2003;327:1032–1037.
10. Aggarwal R, Darzi A. Technical-skills training in the 21st century. *N Engl J Med*. 2006;355:2695–2696.
11. Martin J, Regehr G, Reznick R, et al. Objective structured assessment of techncial skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.
12. Kane M. Validation. In: Brennan R, editor. *Educational Measurement*. 4th ed, Westport, CT: American Council on Education and Praeger; 2006:17–64.
13. Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49:560–575.
14. Hatala R, Cook DA, Brydges R, et al. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Adv Heal Sci Educ [Internet]*. 2015;20:1149–1175.
15. Reznick R, Regehr G, MacRae H, et al. Testing technical skill via an innovative "bench station" examination. *Am J Surg*. 1997;173:226–230.
16. Datta V, Bann S, Mandalia M, et al. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. *Am J Surg*. 2006;192:372–378.
17. White LW, Lendvay TS, Holst D, et al. Using crowd-assessment to support surgical training in the developing world. *J Am Coll Surg [Internet]*. 2014;219:e40.
18. D'Angelo ALD, Rutherford DN, Ray RD, et al. Operative skill: quantifying surgeon's response to tissue properties. *J Surg Res [Internet]*. 2015;198: 294–298.
19. Hu Y-Y, Peyre SE, Arriaga AF, et al. Postgame analysis: using video-based coaching for continuous professional development. *J Am Coll Surg [Internet]*. 2012;214:115–124.
20. Glarner CE, Hu YY, Chen CH, et al. Quantifying technical skills during open operations using video-based motion analysis. *Surgery (United States) [Internet]*. 2014;156:729–734.
21. Watson RA. Use of a machine learning algorithm to classify expertise: analysis of hand motion patterns during a simulated surgical task. *Acad Med [Internet]*. 2014;89:1–5.
22. Mackenzie CF, Watts D, Patel RR, et al. Sensor-free computer vision hand-motion entropy and video analysis of technical performance during open vascular surgery: Proof of concept for methodology. *J Am Coll Surg*. 2016;223:e63. https://doi.org/10.1016/j.jamcollsurg.2016.08.166.
23. Hayter MA, Friedman Z, Bould MD, et al. Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement. *Can J Anesth*. 2009;56:419–426.
24. Datta V, Mackay S, Mandalia M, et al. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg*. 2001;193:479–485.
25. Datta V, Chang A, Mackay S, et al. The relationship between motion analysis and surgical technical assessments. *Am J Surg*. 2002;184:70–73.
26. Reiley CE, Lin HC, Yuh DD, et al. Review of methods for objective surgical skill evaluation. *Surg Endosc Other Interv Tech*. 2011;25:356–366.
27. Chen C-H, Hu YH, Radwin RG. A motion tracking system for hand activity assessment. In 2014 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP) (pp. 320–324). IEEE. 2014. https://doi.org/10.1109/ChinaSIP.2014.6889256.
28. Chen C-H, Hu YH, Yen TY, et al. Automated video exposure assessment of repetitive hand activity level for a load transfer task. *Hum Factors J Hum Factors Ergon Soc [Internet]*. 2012;55:298–308.
29. Chen C-H, Azari DP, Hu YH, et al. The accuracy of conventional 2D video for quantifying upper limb kinematics in repetitive motion occupational tasks. *Ergonomics*. 2015;58:2057–2066. https://doi.org/10.1080/00140139.2015.1051594.
30. Akkas O, Azari DP, Chen C-HE, et al. A hand speed – duty cycle equation for estimating the ACGIH hand activity level rating. *Ergonomics [Internet]*. 2014;58:184–194.
31. Azari DP, Pugh CM, Laufer S, et al. Evaluation of simulated clinical breast exam motion patterns using marker-less video tracking. *Hum Factors [Internet]*. 2015;58:427–440.
32. Frasier LL, Azari DP, Ma Y, et al. A marker-less technique for measuring kinematics in the operating room. *Surgery [Internet]*. 2016;160:1400–1413.
33. Yen TY, Radwin RG. A video-based system for acquiring biomechanical data synchronized with arbitrary events and activities. *IEEE Trans Biomed Eng*. 1995;42:944–948.
34. Annett J. Subjective rating scales: science or art? *Ergonomics*. 2002;45:966–987.
35. Moulton CA, Regehr G, Lingard L, et al. Slowing down to stay out of trouble in the operating room: remaining attentive in automaticity. *Acad Med [Internet]*. 2010;85:1571–1577.
36. D'Angelo ALD, Rutherford DN, Ray RD, et al. Idle time: an underdeveloped performance metric for assessing surgical skill. *Am J Surg [Internet]*. 2015;209:645–651.
37. Pugh CM. Application of national testing standards to simulation-based assessments of clinical palpation skills. *Mil Med [Internet]*. 2013;178(10 suppl):55–63.
38. Laufer S, D'Angelo A-LD, Kwan C, et al. Rescuing the clinical breast examination. *Ann Surg*. 2016;XX:1–6. https://doi.org/10.1097/SLA.0000000000002024.
39. Grober ED, Roberts M, Shin EJ, et al. Intraoperative assessment of technical skills on live patients using economy of hand motion: establishing learning curves of surgical competence. *Am J Surg [Internet]*. 2010;199:81–85.
40. Greiner TM. Hand Anthropometry of U.S. Army Personell. *Tech Rep Natick*. 1991;434. TR-92/011.
41. Hogan N, Sternad D. Sensitivity of smoothness measures to movement duration, amplitude, and arrests. *J Mot Behav [Internet]*. 2009;41:529–534.
42. Rao C, Yilmaz A, Shah M. View-invariant representation and recognition of actions. *Int J Comput Vis*. 2002;50:203–226.
43. Radwin RG, Azari DP, Lindstrom MJ, et al. A frequency–duty cycle equation for the ACGIH hand activity level. *Ergonomics [Internet]*. 2015;58:173–183.
44. Watson RA. Computer-aided feedback of surgical knot tying using optical tracking. *J Surg Educ*. 2012;69:306–310.
45. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr [Internet]*. 1974;19:716–723.
46. Neter J, Wasserman W, Kutner M. Chapters 8–9: Building the Regression Model. In: Hercher RT, Shiell E, eds. *Applied Linear Statistical Models*. 3rd ed., Boston, MA: Richard D. Irwin, INC; 1990. p. 1181.
47. Zumel N, Mount J. *Practical Data Science With R*. 1st ed. Shelter Island, NY: Manning Inc; 2014, 416.
48. Aggarwal R, Grantcharov T, Moorthy K, et al. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg*. 2008;247:372–379.
49. Frischknecht AC, Kasten SJ, Hamstra SJ, et al. The objective assessment of experts' and novices' suturing skills using an image analysis program. *Acad Med*. 2013;88:260–264.
50. Lin HC, Shafran I, Yuh D, et al. Towards automatic skill evaluation: detection and segmentation of robot-assisted surgical motions. *Comput aided Surg*. 2006;11:220–230.