JAMA Network Open

Original Investigation | Health Informatics

Evaluation of Deep Learning Models for Identifying Surgical Actions and Measuring Performance

Shuja Khalid, MSc; Mitchell Goldenberg, MBBS, PhD; Teodor Grantcharov, MD, PhD; Babak Taati, PhD; Frank Rudzicz, PhD

Abstract

IMPORTANCE When evaluating surgeons in the operating room, experienced physicians must rely on live or recorded video to assess the surgeon's technical performance, an approach prone to subjectivity and error. Owing to the large number of surgical procedures performed daily, it is infeasible to review every procedure; therefore, there is a tremendous loss of invaluable performance data that would otherwise be useful for improving surgical safety.

OBJECTIVE To evaluate a framework for assessing surgical video clips by categorizing them based on the surgical step being performed and the level of the surgeon's competence.

DESIGN, SETTING, AND PARTICIPANTS This quality improvement study assessed 103 video clips of 8 surgeons of various levels performing knot tying, suturing, and needle passing from the Johns Hopkins University–Intuitive Surgical Gesture and Skill Assessment Working Set. Data were collected before 2015, and data analysis took place from March to July 2019.

MAIN OUTCOMES AND MEASURES Deep learning models were trained to estimate categorical outputs such as performance level (ie, novice, intermediate, and expert) and surgical actions (ie, knot tying, suturing, and needle passing). The efficacy of these models was measured using precision, recall, and model accuracy.

RESULTS The provided architectures achieved accuracy in surgical action and performance calculation tasks using only video input. The embedding representation had a mean (root mean square error [RMSE]) precision of 1.00 (0) for suturing, 0.99 (0.01) for knot tying, and 0.91 (0.11) for needle passing, resulting in a mean (RMSE) precision of 0.97 (0.01). Its mean (RMSE) recall was 0.94 (0.08) for suturing, 1.00 (0) for knot tying, and 0.99 (0.01) for needle passing, resulting in a mean (RMSE) recall of 0.98 (0.01). It also estimated scores on the Objected Structured Assessment of Technical Skill Global Rating Scale categories, with a mean (RMSE) precision of 0.85 (0.09) for novice level, 0.67 (0.07) for intermediate level, and 0.79 (0.12) for expert level, resulting in a mean (RMSE) precision of 0.77 (0.04). Its mean (RMSE) recall was 0.85 (0.05) for novice level, 0.69 (0.14) for intermediate level, and 0.80 (0.13) for expert level, resulting in a mean (RMSE) recall of 0.78 (0.03).

CONCLUSIONS AND RELEVANCE The proposed models and the accompanying results illustrate that deep machine learning can identify associations in surgical video clips. These are the first steps to creating a feedback mechanism for surgeons that would allow them to learn from their experiences and refine their skills.

JAMA Network Open. 2020;3(3):e201664. doi:10.1001/jamanetworkopen.2020.1664

Key Points

Question Can deep machine learning models be used to assess important surgical characteristics, such as the type of procedure and surgical performance?

Findings In this quality improvement study of 103 video clips of table-top surgical procedures, performed by 8 surgeons and including 4 to 5 trials of 3 surgical actions, deep machine learning obtained a mean precision of 0.97 and a mean recall of 0.98 in detecting surgical actions and a mean precision of 0.77 and a mean recall of 0.78 in estimating the surgical skill level of operators.

Meaning In this study, automatic processing of short surgical video clips by deep machine learning accurately identified and assessed surgical performance.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Introduction

Capturing the most important characteristics of safe surgery has long been the goal of surgical performance evaluation.[1] Several tools for achieving this aim have been proposed. For instance, the Global Operative Assessment of Laparoscopic Skills[2] evaluates the depth perception, bimanual dexterity, efficiency, tissue handling, and autonomy of surgeons, and the Objective Structured Assessment of Technical Skill[3] assesses categories such as respect for tissue, time and motion, instrument handling, knowledge of instruments, flow of operation, use of assistants, and knowledge of the specific procedure. However, these rating tools are not free from bias and can be challenging to implement, given that they require considerable time and effort from experienced surgeons.[4] Despite these concerns, these methods continue to serve as the criterion standards for categorically isolating areas of improvement for surgeons.

In this study, we tested 2 machine learning algorithms to assess surgical performance on labels calibrated across expert raters. The first algorithm transformed video frames into a low-dimensional representation and used deep neural networks to learn the spatiotemporal characteristics of the video. The second approach explicitly captured the pixel-level outlines (ie, a segmentation) of surgical instruments in each frame. Accurately detecting and tracking surgical instruments within each of these clips would allow for a fine-grained analysis of instrument handling, elegance of motion, and autonomy over the course of the surgery. We sought to train our neural networks to estimate surgical actions and performance measures associated with the Objective Structured Assessment of Technical Skill or Global Operative Assessment of Laparoscopic Skills scores based on extracted features from each video. Because the outcomes of interest (eg, dexterity) involve highly interdependent aspects of both spatial and temporal features,[5] we aimed to create neural network models that would be sensitive to the dynamics of change over time.

Methods

This study followed the Standards for Quality Improvement Reporting Excellence (SQUIRE) reporting guideline. This study was approved by the Unity Health Toronto research ethics board.

Data Set

We used the Johns Hopkins University–Intuitive Surgical Gesture and Skill Assessment Working Set,[1] which consists of 103 video clips showing curated table-top surgical setups and includes kinematic measurements (ie, articulation and velocities of joints) from 8 surgeons performing 4 to 5 trials of 3 surgical actions, such as knot tying, needle passing, and suturing.[6] All participants, both patients and surgeons, provided written informed consent. The data were captured using the DaVinci Robotic System (Intuitive Surgical)[7] and came with manually annotated labels that corresponded to performance scores defined by a modified version of Objective Structured Assessment of Technical Skill,[3] specifically the global rating scale (GRS). The GRS excluded certain categories, such as use of assistants, because each clip depicted a surgeon completing a short procedure in a controlled environment where assistance is not available. The GRS used a Likert scale with values ranging from 1 to 5 for respect for tissue, suturing and needle handling, time and motion, flow of operation, overall performance, and quality of product.

The Johns Hopkins University–Intuitive Surgical Gesture and Skill Assessment Working Set has a well-defined validation scheme that allows for a structured comparison of novel algorithms. The scheme includes leave-one-supertrial-out (LOSO), in which 1 of 5 trials is removed from the data set and used for validation for each procedure, and leave-one-user-out (LOUO).[6] This scheme allows for an objective comparison of approaches and has been used by several independent researchers such as DiPietro et al,[8] Sarikaya et al,[9] Lea et al,[5] and Gao et al,[10] who attempted to estimated surgical actions and quantify surgical skill. This data set was collected as a collaboration between Johns

Hopkins University and Intuitive Surgical, within an institutional review board–approved study and has been released for public use.[1]

Sequence Modeling of Embedding Representations

Autoencoders are neural networks that are trained to accurately recreate and therefore represent input data. Specifically, they learn to decompose the input into a smaller set of signals,[11] called an embedding. For example, an autoencoder neural network can consist of 2 stages, as follows: (1) an encoder that compresses information into the embedding and (2) a decoder that tries to reconstruct the original input from the embedding. These 2 stages can be jointly trained to find the optimal, smaller set of dimensions and are depicted in **Figure 1**; the left portion shows the encoder, which applies a series of mathematical operations to the input to decrease data dimensionality, and the right portion shows the decoder, which then applies a series of functions to recreate the original image. The resulting discrepancy in pixel values between the original and reconstructed images is called the reconstruction error, and minimizing this error is the goal of training the network.

In this study, the input and output images were resized to 224 × 224 pixels (with 3 color channels each), and the embeddings consisted of 361 elements, set empirically. Given that the model did not require any additional data or labels for the training process, it is what is known as a self-supervised model.[12] The resulting embeddings represent the frames of the surgical video clips in a much more compact form and are used to train a temporal model. The technical details for implementation have been summarized in the eAppendix in the Supplement.

Sequence Modeling of Key Point Features

Instead of encoding entire video frames as embeddings, we can explicitly represent only the instruments in those frames using key points, which are the pixels on a segment that define that segment (eg, the tip of a needle driver). The presence of a certain instrument as well as its orientation, position, and size are all computable from these key points. We only extracted key points that are required to estimate surgical performance and actions. This method creates a technique called semantic segmentation, which outlines and labels regions in an image,[13] ie, an instrument or the background. Obtaining key point representations requires certain characteristics of the instruments, such as orientation and position, that appear within each frame. We then used a neural network to capture the changes of these characteristics over time.

Statistical Analysis

The statistical tests used to validate the performance of the proposed models were precision, recall, and the F1 score. These metrics are prevalent in the machine learning community for classification tasks. Precision is a measure of the number of true-positives divided by the sum of the true-positives and the false-positives. Recall is a measure of the number of true-positives divided by the sum of true-positives and false-negatives. The F1 score represents the balance that exists between the precision and recall scores. It is defined as the product of precision and recall divided by the sum of precision and recall. All models were trained using PyTorch version 1.3. No prespecified threshold for statistical significance was set. Data were analyzed from March to July 2019.

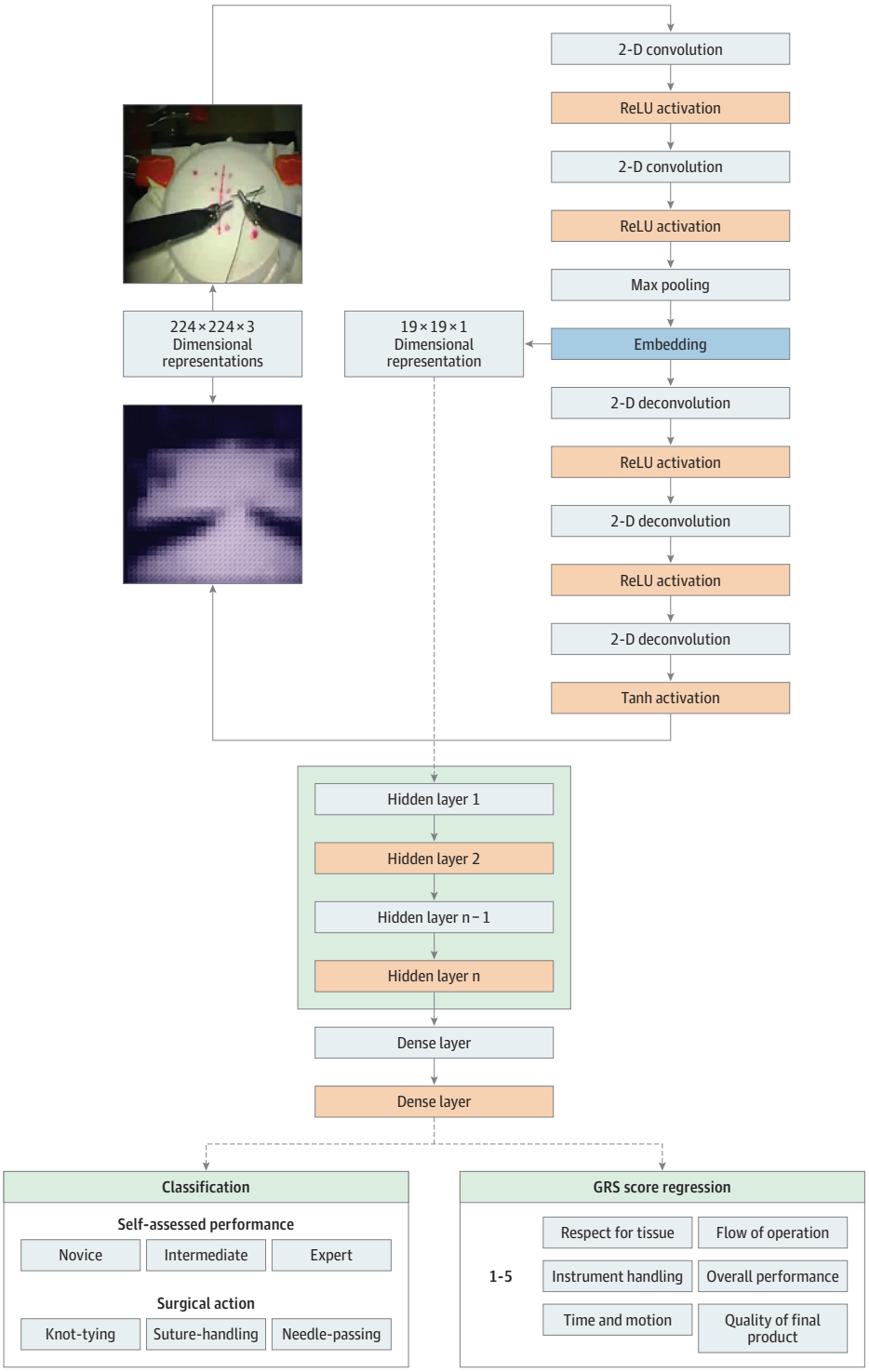
Results

Sequence Modeling of Embedding Representations

We initially experimented with an autoencoder to find the smallest embedding dimension that would allow for constructing a discernible image. Using these embeddings, we trained the autoencoder using both the LOUO and LOSO validation schemes.[6] The LOUO and LOSO schemes require averaging metrics across all 8 surgeons and all 5 trials, respectively. **Table 1**[5,8-10,14-18] and **Table 2**[19] show the results. The embedding representation analysis outperformed the previous state-of-the-art models and did so without using any kinematic data, which was required in previous work that was

not robot assisted. For example, for suturing, the embedding representation analysis using LOSO had a mean (root mean square error [RMSE]) accuracy of 0.97 (0.03), a mean (RMSE) precision of 1.00 (0), a mean (RMSE) recall of 0.94 (0.08), and a mean (RMSE) F1 score of 0.97 (0.04). Using the LOUO, the embedding representation analysis had a mean (RMSE) accuracy of 0.84 (0.20), a mean (RMSE) precision of 1.00 (0), a mean (RMSE) recall of 0.88 (0.21), and a mean (RMSE) F1 score of

Figure 1. Embedding Representation Analysis Architecture



The proposed end-to-end model can be used as a classifier to predict surgical actions and self-reported skill and also as a regression model, which predicts scores on the Likert-scale for each category of the Global Rating Scale (GRS). 2-D indicates two-dimensional; and ReLU, rectified linear unit.

0.92 (0.14). The second highest-performing model on accuracy was from the study by Forestier et al,[17] with an accuracy of 0.94 (RMSE not reported). The second highest mean (RMSE) precision score (0.93 [0.01]), mean (RMSE) recall score (0.93 [0.01]), and mean (RMSE) F1 score (0.92 [0.01]) belonged to the LOSO model presented by Gao et al.[10] Overall, the embedding representation had a mean (RMSE) precision of 1.00 (0) for suturing, 0.99 (0.01) for knot tying, and 0.91 (0.11) for needle passing, resulting in a mean (RMSE) precision of 0.97 (0.01). Its mean (RMSE) recall was 0.94 (0.08) for suturing, 1.00 (0) for knot tying, and 0.99 (0.01) for needle passing, resulting in a mean (RMSE) recall of 0.98 (0.01) (Table 1). Using the LOSO scheme, it estimated scores on the Objected Structured Assessment of Technical Skill Global Rating Scale categories, with a mean (RMSE) precision of 0.85 (0.09) for novice level, 0.67 (0.07) for intermediate level, and 0.79 (0.12) for expert level, resulting in a mean (RMSE) precision of 0.77 (0.04). Its mean (RMSE) recall was 0.85 (0.05) for novice level, 0.69 (0.14) for intermediate level, and 0.80 (0.13) for expert level, resulting in a mean (RMSE) recall of 0.78 (0.03) (Table 2).

Our tool also estimated scores for GRS categories when used as a regression model, with a mean (RMSE) accuracy of 0.54 (0.03) for suture handling, 0.32 (0.14) for time and motion, 0.46 (0.10) for

Table 1. Surgical Task Recognition Literature Review

					Mean (RMSE)		
Source	Data type	Scheme	Model	Metric	ST	KT	NP
Gurcan et al[14]	Kinematic	Holdout	MS-RNN	Accuracy	0.90 (0.08)	0.90 (0.08)	0.90 (0.08)
Sarikaya et al[9]	Kinematic	Holdout	Optical flow	Accuracy	0.91 (0.01)	0.88 (0.03)	0.74 (0.04)
Tao et al[15]	Kinematic	LOSO	Sparse HMM	Accuracy	0.81 (NA)	0.76 (NA)	0.83 (NA)
		LOUO	Sparse HMM	Accuracy	0.68 (NA)	0.59 (NA)	0.66 (NA)
DiPietro et al[8]	Kinematic	LOUO	bi-LSTM	Accuracy	0.83 (0.06)	0.83 (0.06)	0.83 (0.06)
Sefati et al[16]	Kinematic	LOUO	SC-CRF	Accuracy	0.80 (NA)	0.79 (NA)	0.75 (NA)
Forestier et al[17]	Kinematic	LOSO	DIP	Accuracy	0.94 (NA)	0.93 (NA)	0.81 (NA)
		LOUO	DIP	Accuracy	0.88 (NA)	0.90 (NA)	0.75 (NA)
Gao et al[10]	Kinematic	LOSO	AS-DTW	Precision	0.93 (0.01)	0.93 (0.01)	0.93 (0.01)
				Recall	0.93 (0.01)	0.93 (0.01)	0.93 (0.01)
				F1 score	0.92 (0.01)	0.92 (0.01)	0.92 (0.01)
		LOUO	AS-DTW	Precision	0.91 (0.01)	0.93 (0.01)	0.91 (0.01)
				Recall	0.90 (0.01)	0.90 (0.01)	0.90 (0.01)
				F1 score	0.89 (0.01)	0.89 (0.01)	0.89 (0.01)
Lea et al[5]	Video	LOUO	ST-CNN with segmentation	Accuracy	0.74 (NA)	0.74 (NA)	0.74 (NA)
Tao et al[15]	Video	LOSO	BoSTF	Accuracy	0.85 (NA)	0.72 (NA)	0.84 (NA)
		LOUO	BoSTF	Accuracy	0.76 (NA)	0.62 (NA)	0.79 (NA)
Liu et al[18]	Video	LOUO	TCN	Accuracy	0.82 (NA)	0.82 (NA)	0.82 (NA)
Embedding representation analysis	Video	LOSO	bi-LSTM (attention)	Accuracy	0.97 (0.03)	0.97 (0.03)	0.97 (0.03)
				Precision	1.00 (0)	0.99 (0.01)	0.91 (0.11)
				Recall	0.94 (0.08)	1.00 (0)	0.99 (0.01)
				F1 score	0.97 (0.04)	0.99 (0.01)	0.95 (0.06)
		LOUO	bi-LSTM (attention)	Accuracy	0.84 (0.20)	0.84 (0.20)	0.84 (0.20)
				Precision	1.00 (0)	0.92 (0.22)	0.75 (0.43)
				Recall	0.88 (0.21)	1.00 (0)	0.70 (0.42)
				F1 score	0.92 (0.14)	0.94 (0.17)	0.72 (0.42)
Key point representation analysis	Video	LOSO	bi-LSTM	Accuracy	0.36 (0.04)	0.36 (0.04)	0.36 (0.04)
				Precision	1.00 (0.04)	0.01 (0.04)	0.01 (0.01)
				Recall	1.00 (0.25)	0.35 (0.17)	0.32 (0.13)
				F1 score	1.00 (0.18)	0.13 (0.07)	0.03 (0.01)

Abbreviations: AS-DTW, asymmetric subsequence–dynamic time warping; bi-LSTM, bidirectional long short-term memory; BoSTF, bag of spatiotemporal features; DIP, discriminative interpretable patterns; HMM, hidden Markov model; KT, knot tying; LOSO, leave-one-supertrial-out; LOUO, leave-one-user-out; MS-RNN, multimodal stochastic recurrent neural network; NA, not applicable; NP, needle passing; RMSE, root mean square error; SC-CRF, skip chain–conditional random field; ST, suturing; ST-CNN, spatiotemporal convolutional neural network; TCN, temporal convolutional network.

flow of operation, 0.41 (0.12) for overall performance, and 0.51 (0.10) for quality of final product. The architecture of the tool as a regression model is visualized in Figure 1.

Sequence Modeling Key-Point Feature Representations

To create a representation of an instrument using key points, we performed a preliminary subjective analysis for context. If automatically detecting key points and labeling segments were not sufficiently accurate, the temporal classifier meant to identify the dynamics of the surgical procedure would be negatively affected.

Examples of correct and incorrect segmentations are shown in **Figure 2**. Each image contains the original image overlaid with the segmentations created by the neural network. It is important to also consider the cases where either the segmentations or the associated class were incorrectly chosen.

Based on the qualitative analysis of frame-level segmentations in Figure 2, the results for this method were limited because of the inability of the segmentation model to yield consistent results across the constituent frames of each surgical clip. The estimated scores for the GRS categories demonstrate this; the per-category mean (RMSE) validation accuracies fluctuated between 0.32 (0.14) for time and motion to 0.54 (0.03) for suture handling.

Discussion

In this study, modeling sequences with neural network embeddings provided state-of-the-art results in surgical action detection using only video input. The traditional key point representation, which use explicit representations of the instruments, was highly sensitive to the preliminary segmentation of the instruments and may not generalize well. Several studies have combined video and kinematic data to evaluate performance and recognize actions. For example, a study by Jin et al[20] quantified operative skill using a deep neural network called Fast R-CNN[21] with a subset of frames from the Modeling and Monitoring of Computer Assisted Interventions 2016 tools data set[22] to detect surgical instruments in each frame. This neural network directly analyzed aspects of texture in the image to estimate the location of instruments as well as their trajectories, movement range, and economy of motion. On a set of 4 test videos, they distinguished between experienced and inexperienced

Table 2. Surgical Skill Assessment Literature Review in Terms of Accuracy, Precision, Recall, and F1 Score, Given Kinematic or Video Input and Novice, Intermediate, and Expert Skill Levels

					Surgical skill level, mean (RMSE)		
Source	Data type	Scheme	Model	Metric	Novice	Intermediate	Expert
Wang et al[19]	Kinematic	LOSO	CNN	Accuracy	0.93 (NA)	0.89 (NA)	0.85 (NA)
				F1 score	0.94 (NA)	0.75 (NA)	0.93 (NA)
		Holdout	CNN	F1 score	0.95 (NA)	0.77 (NA)	0.94 (NA)
Embedding analysis	Video	LOSO	bi-LSTM (attention)	Accuracy	0.77 (0.14)	0.77 (0.14)	0.77 (0.14)
				Precision	0.85 (0.09)	0.67 (0.07)	0.79 (0.12)
				Recall	0.85 (0.05)	0.69 (0.14)	0.80 (0.13)
				F1 score	0.85 (0.07)	0.68 (0.10)	0.79 (0.12)
		LOUO	Gated recurrent unit	Accuracy	0.70 (0.21)	0.70 (0.21)	0.70 (0.21)
				Precision	0.91 (0.05)	0.48 (0.11)	0.70 (0.15)
				Recall	0.76 (0.08)	0.67 (0.19)	0.75 (0.11)
				F1 score	0.83 (0.05)	0.55 (0.12)	0.72 (0.13)
Key point representation analysis	Video	LOUO	bi-LSTM	Accuracy	0.73 (0.33)	0.73 (0.33)	0.73 (0.33)
				Precision	1.00 (0)	0.01 (0)	1.00 (0)
				Recall	0.47 (0.18)	0.29 (0.11)	1.00 (0)
				F1 score	0.64 (0.25)	0.02 (0.01)	1.00 (0)

Abbreviations: bi-LSTM, bidirectional long short-term memory; CNN, convolutional neural network; LOSO, leave-one-supertrial-out; LOUO, leave-one-user-out; NA, not applicable; RMSE, root mean square error.

surgeons. In contrast, we extracted more information across time, and instead of placing bounding boxes around regions of interest,[20] we segmented each instrument at the pixel level, which allowed us to analyze orientation.

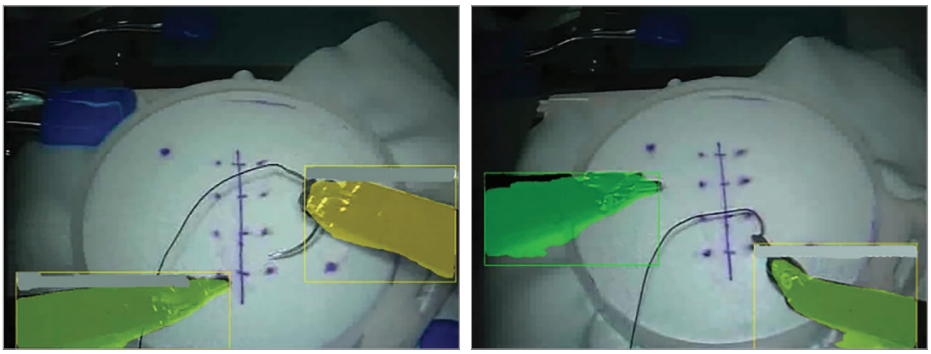
Skill assessment has been widely explored through analyzing kinematic data in robot-assisted surgeries. This type of data allows for classic machine learning methods to learn the inherent structure of the data. Wang and Fey[19,23] and Hung et al[24] used kinematic data acquired from the OnSite computer built into the DaVinci Robotic System[7] and used neural networks to classify performance and, in the study by Hung et al,[24] associated the results with patient outcomes, such as surgery time, estimated blood loss, length of stay, pelvic drainage volume, drainage tube duration, and Foley duration. DiPietro et al,[8] Sefati et al,[16] Forestier et al,[17] and Gao et al[10] leveraged a variety of classic and novel methods that used kinematic data to predict surgical gestures. These approaches quantified surgical skill assessment by applying a variety of traditional machine learning techniques to kinematic data exclusive to robot-assisted surgery. In contrast, our end-to-end model can estimate surgical performance directly from raw video, including standardized rating scores, such as the GRS.[6]

Other studies by Twinanda et al,[25] Tao et al,[15] and Sahu et al[26] inferred temporal associations from video and kinematics to detect surgical phase. These methods used video data to augment the available kinematic data and provide the models with additional context during training. The results improved on previous methods, which can be seen in Table 1.

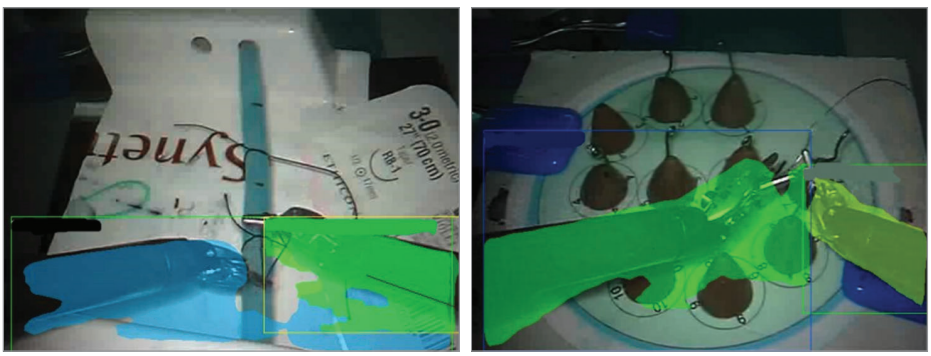
Liu et al[18] proposed another technique for gesture segmentation, using temporally and spatially coherent features from surgical clips. Sarikaya et al[9] explored deep convolutional optical flow models for gesture recognition and claimed competitive results. These studies used video data to detect surgical phase, in contrast to the other approaches. This deep learning approach to extract visual cues directly from videos has been shown to provide competitive results. We not only exceeded the

Figure 2. Predicted Segmentations Overlaid on Original Image

A Examples of correctly segmented instruments



B Examples of incorrectly segmented instruments



performance of these models in surgical action detection using visual cues but also extended this approach to evaluating performance.

Therefore, our methods would generalize to surgical procedures without robotic assistance and are the first to use purely visual cues to estimate validated performance scores and GRS[6] scores directly from surgical video. Our results provide a framework to automatically evaluate performance during surgical tasks, which has the potential to provide feedback to surgeons, potentially in the context of effective curriculum creation and advanced surgical education.

Limitations

This study has limitations. In deep learning, large data sets are often required to sufficiently train models that can generalize to real world scenarios. Despite the utility of the Johns Hopkins University–Intuitive Surgical Gesture and Skill Assessment Working Set, larger data sets will be required to extensively test the proposed architectures. These data sets would ideally include a larger variety of procedures from more surgeons and would be recorded in actual surgical settings. The variability and subjectivity in the coding of coarse GRS performance scores can also make performance estimation challenging. If the labels are biased in any way, then they will infect the model trained from them.

Conclusions

In this study, our proposed neural network models inferred temporal patterns from surgical instrument motions and associated them with surgical gestures, actions, and performance-related cues given validated rating scales. The models achieved state-of-the-art results in both surgical action recognition and performance recognition, requiring only video data. These machine learning approaches obtained a mean precision of 91% and mean recall of 94% in detecting surgical actions, and a mean precision of 77% and mean recall of 78% in predicting the surgical skill level of operators. The use of video data alone for these predictions generalizes to other types of surgery, for which surgical robotic sensory data are not available.

ARTICLE INFORMATION

Accepted for Publication: February 3, 2020.

Published: March 30, 2020. doi:10.1001/jamanetworkopen.2020.1664

Corresponding Author: Frank Rudzicz, PhD, Surgical Safety Technologies, 20 Queen St W, 35th Floor, Toronto, ON M5H 3R3, Canada (f.rudzicz@surgicalsafety.com).

Author Affiliations: Surgical Safety Technologies, Toronto, Ontario, Canada.

Author Contributions: Mr Khalid and Dr Rudzicz had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: All authors.

Acquisition, analysis, or interpretation of data: Khalid.

Drafting of the manuscript: Khalid, Goldenberg, Rudzicz.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Khalid.

Obtained funding: Grantcharov.

Administrative, technical, or material support: Goldenberg, Grantcharov.

Supervision: Goldenberg, Grantcharov, Taati, Rudzicz.

Conflict of Interest Disclosures: Mr Khalid and Drs Goldenberg, Grantcharov, Taati, and Rudzicz reported having a patent pending related to measuring surgical performance using deep learning with Surgical Safety Technologies.

Mr Khalid reported receiving personal fees from Surgical Safety Technologies during the conduct of the study and outside the submitted work. Dr Goldenberg reported receiving personal fees from Surgical Safety Technologies during the conduct of the study. Dr Taati reported receiving personal fees from Surgical Safety Technologies during the conduct of the study and outside the submitted work. Dr Rudzicz reported receiving salary from Surgical Safety Technologies during the conduct of the study and outside the submitted work.

Funding/Support: The study was supported by Surgical Safety Technologies.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data and preparation, review, or approval of the manuscript. The company approved the decision to submit the manuscript for publication.

REFERENCES

1. Niitsu H, Hirabayashi N, Yoshimitsu M, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg Today*. 2013;43(3):271-275. doi:10.1007/s00595-012-0313-7

2. 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. doi:10.1016/j.amjsurg.2005.04.004

3. Martin JA, Regehr G, Reznick R, et al. Objective Structured Assessment of Technical Skill (OSATS) for surgical residents. *Br J Surg*. 1997;84(2):273-278. doi:10.1002/bjs.1800840237

4. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Adv Health Sci Educ Theory Pract*. 2015;20(5):1149-1175. doi:10.1007/s10459-015-9593-1

5. Lea C, Reiter A, Vidal R, Hager GD. Segmental spatiotemporal CNNs for fine-grained action segmentation. Accessed February 25, 2020. https://arxiv.org/abs/1602.02995

6. Gao Y, Swaroop Vedula S, Reiley CE, et al. JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS): a surgical activity dataset for human motion modeling. Accessed February 25, 2020. https://cirl.lcsr.jhu.edu/wp-content/uploads/2015/11/JIGSAWS.pdf

7. Intuitive Surgical. DaVinci by Intuitive. Accessed February 25, 2020. https://www.intuitive.com/en-us/products-and-services/da-vinci

8. DiPietro R, Lea C, Malpani A, et al. Recognizing surgical activities with recurrent neural networks. Accessed February 25, 2020. https://arxiv.org/abs/1606.06329

9. Sarikaya D, Jannin P. Surgical gesture recognition with optical flow only. Accessed February 25, 2020. https://arxiv.org/abs/1904.01143

10. Gao Y, Vedula SS, Lee GI, Lee MR, Khudanpur S, Hager GD. Query-by-example surgical activity detection. *Int J Comput Assist Radiol Surg*. 2016;11(6):987-996. doi:10.1007/s11548-016-1386-3

11. Vincent P, Larochelle H, Bengio Y, Manzagol P-A. Extracting and composing robust features with denoising autoencoders. Accessed February 25, 2020. https://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf

12. Tung H-YF, Tung H-W, Yumer E, Fragkiadaki K. Self-supervised learning of motion capture. Accessed February 25, 2020. https://arxiv.org/abs/1712.01337

13. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. Accessed February 25, 2020. https://arxiv.org/abs/1703.06870

14. Gurcan I, Nguyen HV. Surgical activities recognition using multi-scale recurrent networks. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2019:2887–2891. doi:10.1109/ICASSP.2019.8683849

15. Tao L, Zappella L, Hager GD, Vidal R. Surgical gesture segmentation and recognition. *Med Image Comput Assist Interv*. 2013;16(Pt 3):339-346.

16. Sefati S, Cowan NJ, Vidal R. Learning shared, discriminative dictionaries for surgical gesture segmentation and classification. Accessed February 25, 2020. http://www.vision.jhu.edu/assets/SefatiM2CAI15.pdf

17. Forestier G, Petitjean F, Senin P, et al. Surgical motion analysis using discriminative interpretable patterns. *Artif Intell Med*. 2018;91:3-11. doi:10.1016/j.artmed.2018.08.002

18. Liu D, Jiang T. Deep reinforcement learning for surgical gesture segmentation and classification. Accessed February 25, 2020. https://arxiv.org/abs/1806.08089

19. Wang Z, Fey AM. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. Accessed February 25, 2020. https://arxiv.org/abs/1806.05796

20. Jin A, Yeung S, Jopling J, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. Accessed February 25, 2020. https://arxiv.org/abs/1802.08774

21. Girshick R. Fast R-CNN. Accessed February 25, 2020. https://arxiv.org/abs/1504.08083

22. Medical Image Computing and Computer Assisted Intervention Society. Tool presence detection challenge results. Accessed February 25, 2020. http://camma.u-strasbg.fr/ m2cai2016/index.php/tool-presence-detection-challenge-results

23. Wang Z, Fey AM. SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. Accessed February 25, 2020. https://arxiv.org/abs/1806.05798

24. Hung AJ, Chen J, Che Z, et al. Utilizing machine learning and automated performance metrics to evaluate robot-assisted radical prostatectomy performance and predict outcomes. *J Endourol*. 2018;32(5):438-444. doi: 10.1089/end.2018.0035

25. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. Accessed February 25, 2020. https://arxiv.org/abs/1602.03012

26. Sahu M, Mukhopadhyay A, Szengel A, Zachow S. Tool and phase recognition using contextual CNN features. Accessed February 25, 2020. https://arxiv.org/abs/1610.08854

SUPPLEMENT.

eAppendix. Implementation Details