International Journal of Computer Assisted Radiology and Surgery (2018) 13:1959–1970
https://doi.org/10.1007/s11548-018-1860-1

ORIGINAL ARTICLE

Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery

Ziheng Wang[1] · **Ann Majewicz Fey**[1,2]

Received: 11 January 2018 / Accepted: 11 September 2018 / Published online: 25 September 2018

Abstract

Purpose With the advent of robot-assisted surgery, the role of data-driven approaches to integrate statistics and machine learning is growing rapidly with prominent interests in objective surgical skill assessment. However, most existing work requires translating robot motion kinematics into intermediate features or gesture segments that are expensive to extract, lack efficiency, and require significant domain-specific knowledge.

Methods We propose an analytical deep learning framework for skill assessment in surgical training. A deep convolutional neural network is implemented to map multivariate time series data of the motion kinematics to individual skill levels.

Results We perform experiments on the public minimally invasive surgical robotic dataset, JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS). Our proposed learning model achieved competitive accuracies of 92.5%, 95.4%, and 91.3%, in the standard training tasks: *Suturing*, *Needle-passing*, and *Knot-tying*, respectively. Without the need of engineered features or carefully tuned gesture segmentation, our model can successfully decode skill information from raw motion profiles via end-to-end learning. Meanwhile, the proposed model is able to reliably interpret skills within a 1–3 second window, without needing an observation of entire training trial.

Conclusion This study highlights the potential of deep architectures for efficient online skill assessment in modern surgical training.

Keywords Surgical robotics · Surgical skill evaluation · Motion analysis · Deep learning · Convolutional neural network

Introduction

Due to the prominent demand for both quality and safety in surgery, it is essential for surgeon trainees to achieve required proficiency levels before operating on patients [1]. An absence of adequate training can significantly compromise the clinical outcome, which has been shown in numerous studies [2–4]. Effective training and reliable methods to assess surgical skills are thus critical in supporting trainees in technical skill acquisition [5]. Simultaneously, current surgical training is undergoing significant changes with a rapid uptake of minimally invasive robot-assisted surgery. However, despite advances of surgical technology, most assessments of trainee skills are still performed via outcome-based analysis [6], structured checklists, and rating scales [7–9]. Such assessment requires large amounts of expert monitoring and manual ratings, and can be inconsistent due to biases in human interpretations [10]. Considering the increasing attention to the efficiency and effectiveness of assessment and targeted feedback, conventional methods are no longer adequate in advanced surgery settings [11].

Modern robot-assisted surgical systems are able to collect a large amount of sensory data from surgical robots or simulators [12]. This high volume data could reveal valuable information related to the skills and proficiencies of the operator. However, analyzing such complex surgical data can be challenging. Specifically, surgical motion profiles, by nature, are nonlinear, non-stationary stochastic processes [13,14] with large variability, both throughout a procedure, as well within repetitions of the same type of surgical task (e.g., suture throws) [15]. In addition, the high dimension-

✉ Ziheng Wang
zihengwang@utdallas.edu

[1] Department of Mechanical Engineering, University of Texas at Dallas, Richardson, TX 75080, USA

[2] Department of Surgery, UT Southwestern Medical Center, Dallas, TX 75390, USA

ality of the data creates an additional challenge for accurate and robust skill assessments [10]. Further, although several surgical assessment methods have been developed, methods to autonomously coach the trainee are lacking. Toward this aim, there is a great need to develop techniques for quicker and more effective surgical skill acquisition [11,16]. In this paper, we are particularly interested in online skill assessment methods that could pave the way for autonomous surgical coaching.

Previous approaches in objective skill assessment

Different objective skill assessment techniques have been reported in the literature [16]. Current approaches with a focus on surgical motions can be divided into two main categories: (1) descriptive statistic analysis, and (2) predictive modeling-based methods. Descriptive statistic analysis aims to compute features from motion observations to quantitatively describe skill levels. Specifically, summary features, such as movement time [17–19], path length [17], motion jerk [18], and curvature [17], are widely used and have shown to have high correlations with surgical skills. Other novel measures of motion, such as energy expenditure [20], semantic labels [21], tool orientation [22], and force [19], can also provide discriminative information in measuring skills. However, this approach involves manual feature engineering, requiring task-specific knowledge and significant effort to design optimal skill metrics [23]. In fact, defining the best metrics to capture adequate information and be generalized enough to apply across different types of surgery or groups of surgeons remains an open problem [16,17,24,25].

In contrast to descriptive analysis, predictive modeling-based methods aim to predict surgical skills from motion data. This method can be further categorized into (1) descriptive, and (2) generative modeling. In descriptive modeling, models are learnt by transforming raw motion data to intermediate interpretations and summary features. Coupled with advanced feature selection, these predefined representations are subsequently fed into learning models as an input for skill assessment. In the literature, machine learning (ML) algorithms are commonly explored for modeling, such as k-nearest neighbors (kNN), logistic regression (LR), support vector machines (SVM), and linear discriminant analysis (LDA). Such algorithms yielded a skill predictive accuracy between 61.1 and 95.2% [24,26–28]. Forestier et al. developed a novel vector space model (VSM) to assess skills via learning from the *bag of word*, a collection of discretized local features (strings) obtained from motion data [29]. In Brown et al. [30], explored an ensemble approach, which combines multiple ML algorithms for modeling, and was able to predict rating scores with moderate accuracies (51.7–75.0%). More recently, Zia et al. utilized nearest neighbor (NN) classifiers with a novel feature fusion (texture-, frequency- and entropy-based features) and further improved skill assessment with accuracy ranging from 99.7 to 100% [31]. Although the descriptive modeling-based approaches show their validity in revealing skill patterns and underlying operation structures, the model accuracy and validity are typically limited by the quality of extracted features. Considering the complex nature of surgical motion profiles, critical information has the potential to be discarded within the feature extraction and selection process. Alternatively, in generative modeling, temporal motion data are usually segmented into a series of predefined rudimentary gestures for certain surgical tasks. Using generative modeling algorithms, such as Hidden Markov Model (HMM) and its variants, several class-specific skill models were trained for each level and achieved accuracy ranging from 94.4 to 100% [15,32]. However, the segmentation of surgical gestures from surgeon motions can be a strenuous process. HMM models usually require large amounts of time and computational effort for parameter tuning and model development. Further, one typical deficiency is that the skill assessment is obtained at the global task level, i.e., at the end of each operation. It requires an entire observation for each trial. This drawback potentially undermines the goal of an efficient online surgical skill assessment.

Proposed approach

Deep learning, also referred to as deep structured learning, is a set of learning methods that allow a machine to automatically process and learn from input data via hierarchical layers from low to high levels [33,34]. These algorithms perform feature self-learning to progressively discover abstract representations during the training process. Due to its superiority in complex pattern recognition, this approach dramatically improves the state of the art. Currently, deep learning models have achieved success in strategic games [35], speech recognition [36], medical imaging [37], health informatics [38], and more. In the study of robotic surgical training, DiPietro et al. first apply deep learning based on recurrent neural networks for gesture and high-level task recognition [39]. Still, relatively little work has been done to explore deep learning approaches for surgical skill assessment.

In this paper, we introduce and evaluate the applicability of deep learning for a proficient surgical skill assessment. Specifically, a novel analytical framework with deep surgical skill model is proposed to directly process multivariate time series via an automatic learning. We hypothesize that the learning-based approach could help to explore the intrinsic motion characteristics for decoding skills and promote an optimal performance in online skill assessment systems. Figure 1 shows the end-to-end pipeline framework. Without performing manual feature extraction and selection, latent feature learning is automatically employed on multivariate motion data and directly outputs classifications. To validate

our approach, we conduct experiments on the public robotic surgery dataset, JIGSAW [40], in analysis of three independent training tasks: *Suturing* (SU), *Needle-passing* (NP), and *Knot-tying* (KT). To the best of our knowledge, it is the first study to employ a deep architecture for an objective surgical skill analysis. The main contributions of this paper can be summarized as:

- An novel end-to-end analytical framework with deep learning for skill assessment based on high-level analysis of surgical motion.
- Experimental evaluation of our proposed deep skill model.
- Application of data augmentation leveraging the limitation of small-scale JIGSAWS dataset, discussion on the effect of labeling approaches on the assessment accuracy, and exploration of validation schemes applicable for deep-learning-based development.

In the remainder of this paper we first present our proposed approach and implementation details in "Deep surgical skill classification model" section. We then conduct experiments on JIGSAW dataset to validate the model in "Experiment setup" section. Data preprocessing, training, and evaluation approaches are given. Then, we present our results in "Results" section and discussions in "Discussions" section. Last, we conclude this paper in "Conclusion" section.

Deep surgical skill classification model

Our deep learning model for surgical skill assessment is motivated from studies in multiple domains [34,41,42]. In this section, we introduce a deep architecture using convolutional neural network (CNN) to assess surgical skills from an end-to-end classification.

Problem formulation

Here, the assessment of surgical skills is formalized as a supervised three-class classification problem, where the input is multivariate time series (MTS) of motion kinematics measured from surgical robot end-effectors, X, and the output is the predicted labels representing corresponding expertise levels of trainees, which can be one-hot encoded as $y \in \{1 : \text{“}Novice\text{”}, 2: \text{“}Intermediate\text{”}, 3 : \text{“}Expert\text{”}\}$. Typically, ground-truth skill labels are acquired from expert ratings, crowdsourcing, or self-reporting experience. The objective cost function for training the network is defined as a multinomial cross-entropy cost, J, as shown in Eq. 1.

$$J(\theta) = -\sum_{i=1}^{m}\sum_{k=1}^{K} 1\{y^{(i)} = k\} \log p(y^{(i)} = k|x^{(i)}; \theta) \quad (1)$$

where m is the total number of training examples, K is the class number, $K = 3$, and $p(y^{(i)} = k|x^{(i)}; \theta)$ is the conditional likelihood that the predicted label $y^{(i)}$ on a single training example $x^{(i)}$ is assigned to class $k \in K$, given specific trained model parameters θ.

Model architecture

The architecture of the proposed neural network consists of five types of layers: convolutional layer, pooling layer, flatten layer, fully connected layer and softmax layer. Figure 2 shows a 10-layer working architecture and parameter settings used in the network. Note that the depth of the network is chosen after trial-and-error from the training/validation procedure.

The network takes the slide of length W from C-channel sensory measurements as input, which is a $W \times C$ matrix, where C is the number of channels, or dimensions, of the input time series. Then, input samples are first processed by three convolution–pooling (Conv–pool) stages, where each stage consists of a convolution layer and a max-pooling layer. Each convolution layer has different numbers of kernels with the size of 2 and each kernel is convoluted with the input matrix of the layer with a stride of 1. Specifically, the first convolution (*Conv1*) filters the $W \times 38$ input matrix with 38 kernels; the second convolution with 76 kernels (*Conv2*) will filter the corresponding output matrix of previous layer; and the third convolutional layer (*Conv3*) filters with 152 kernels. To reduce the dimensionality of the feature maps and avoid overfitting, corresponding connections of each convolution are followed by a max-pooling operation. The max-pooling operations take the output of convolution layer as input, and downsample the extracted feature maps, where each local input patch is replaced by taking the maximum value of each channel over the patch. The size of max-pooling is set as 2 with a stride of 2. In this network, we use the rectified linear unit (ReLU) as the activation function to add nonlinearity in all convolutional layers and fully connected layers [43]. Finally, we apply a softmax logistic regression to produce a distribution of probability over three classes for the output layer.

Implementation

To implement the proposed architecture, the deep learning skill model is trained from scratch, which does not require any pre-trained model. The network algorithm is implemented using Keras library with Tensorflow backend based on Python 3.6 [44]. We first initialize parameters at each layer using the Xavier initialization method [34], where biases are

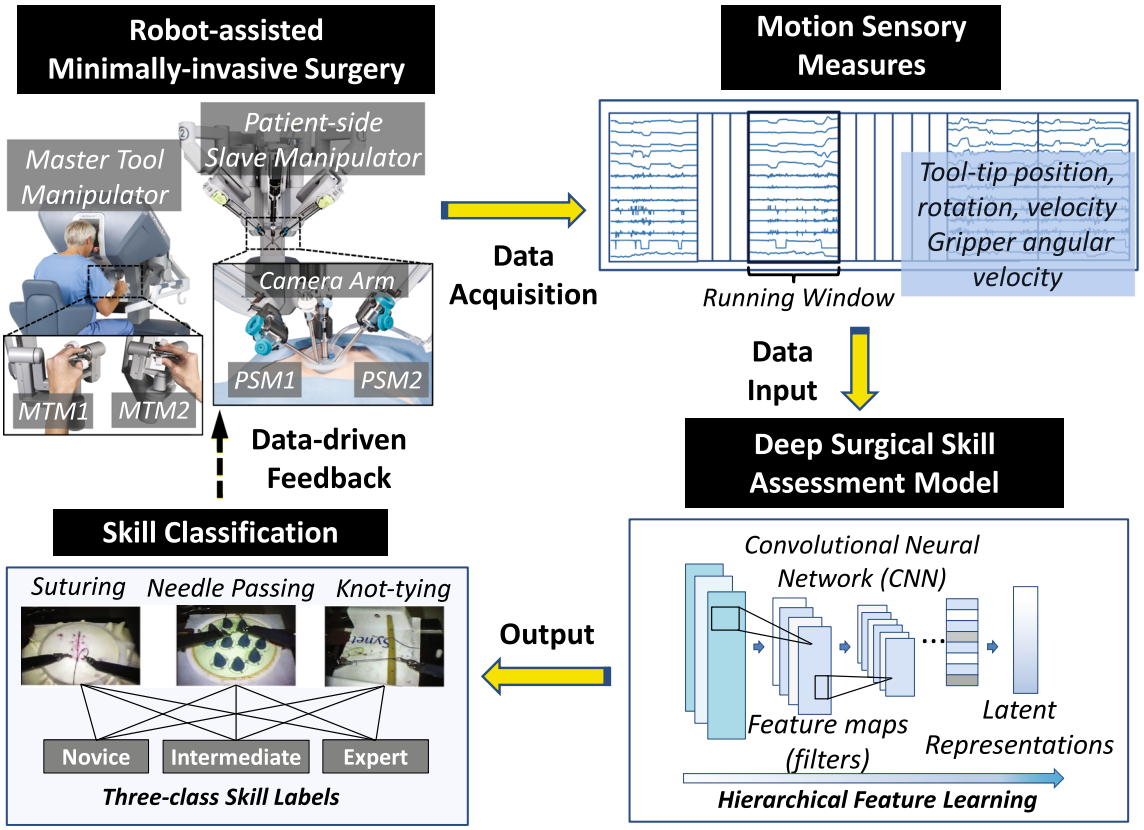

Fig. 1 An end-to-end framework for online skill assessment in robot-assisted minimally invasive surgery. The framework utilizes window sequences of multivariate motion data as an input, recorded from robot end-effectors, and outputs a discriminative assessment of surgical skills via a deep learning architecture

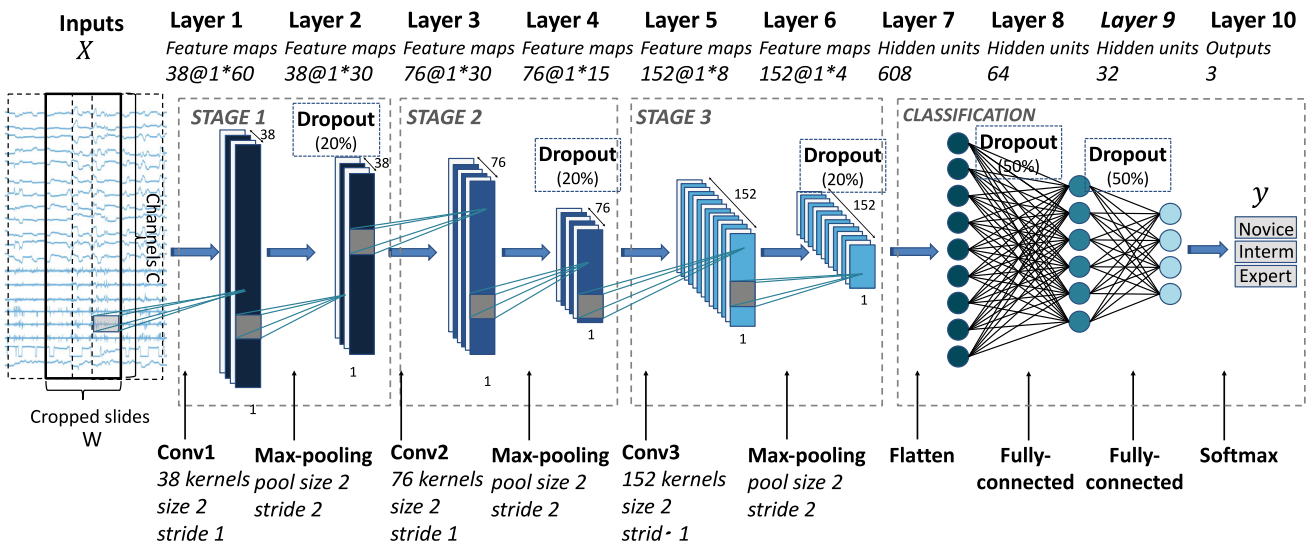

Fig. 2 Illustrations of the proposed deep architecture using a 10-layer convolutional neural network. The window width W used in this example is 60. Starting from the inputs, this model consists of three conv–pool stages with a convolution and max-pooling each, one flatten layer, two fully connected layers, and one softmax layer for outputs. Note that the max-pooling dropout (with probability of 20%) and fully connected dropout (with probability of 50%) is applied during training

initialized as zeros, the weights at each layer are initialized from a Gaussian distribution with mean 0 and a variance of $1/N$, where N specifies the number of neurons in the previous layer.

During the optimization, our network is trained end-to-end by minimizing the multinomial cross-entropy cost between the predicted and ground-truth labels, as defined in Eq. 2, at the learning rate, ε, of 0.0001. To train the net efficiently, we run mini-batch updates of gradient descent, which calculate network parameters on a subset of the training data at each iteration [45]. The size of mini batches is set to 600. A total of 300 epochs for training were run in this work. The network parameters are optimized by an Adam solver [46], which computes adaptive learning rates for each neuron parameter via estimates of first and second moment of the gradients. The exponential decay rates of the first and second moment estimates are set to 0.9 and 0.999, respectively. Also, to achieve better generalization and model performance, we apply a stochastic dropout regularization to our neural network during training time. Components of outputs from specific layers of networks are randomly dropped out at a specific probability [47]. This method has proven its effectiveness to reduce overfitting in complex deep learning models [48]. In this study, we implement two strategies of dropout: one is the max-pooling dropout on the layers of max-pooling after ReLU nonlinearity; another regularization is the fully connected dropout on the fully connected layers. The probabilities of dropout for the max-pooling and fully connected dropout are set at 0.2 and 0.5, respectively. As mentioned above, the hyper-parameters used for CNN implementation include the learning rate, mini-batch size, epoch, number of filters, stride and size of kernel, and dropout rates in the max-pooling and fully connected layers. These hyper-parameters are chosen and fine-tuned by employing the validation set, which is split from training data. We save the best model, as evaluated on validation data, in order to obtain an optimal prediction performance.

Experiment setup

Dataset

Our dataset comes from the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS), the only public-available minimally invasive surgical database, which is collected from the *da Vinci* tele-robotic surgical system [40,49].

The *da Vinci* robot is comprised of two master tool manipulators (MTMs) on left and right sides, two patient-side slave manipulators (PSMs), and an endoscopic camera arm. Robot motion data are captured (sampling frequency 30 Hz) as multivariate time series with 19 measurements for each end-effector: tool tip Cartesian positions (x, y, z), rotations (denoted by a 3×3 matrix R), linear velocities (v_x, v_y, v_z), angular velocities (ω'_x, ω'_y, ω'_z), and the gripper angle θ. Details of the JIGSAWS kinematic motion data are summarized in Table 1.

The dataset contains recordings from eight surgeons with varying robotic surgical experience. Each surgeon performed three different training tasks, namely *Suturing* (SU), *Knot-tying* (KT), and *Needle-passing* (NP), and repeated each task five times. All three tasks are typically standard components in surgical skill training curricula [40]. An illustration of the three operation tasks is shown in Fig. 3. The two ways in which skill labels are reported in JIGSAWS dataset are: (1) self-proclaimed skill labels based on practice hours with *expert* reporting greater than 100 h, *intermediate* between 10 and 100 h, and *novice* reporting less than 10 h of total surgical robotic operation time, and (2) a modified global rating scale (GRS) ranging from 6 and 30, manually graded by an experienced surgeon. In this study, we use the self-proclaimed skill levels and GRS-based skill levels as the ground-truth labels for each surgical trial, respectively. In order to label surgeons skill levels using GRS scores, inspired from [24], thresholds of 15 and 20 are used to divide surgeons into *novice*, *intermediate*, and *expert*, in tasks of *Needle-passing* and *Knot-tying*, and thresholds of 19 and 24 are used in *Suturing* for skill labeling.

Data preparation & inputs

Z-normalization Due to differences in the scaling ranges and offset effects of the sensory data, the data fed into the neural network are first normalized with a z-normalization process. Each channel of raw data, x, is normalized individually as $z = \frac{x-\mu}{\sigma}$, where μ and α are the mean and standard deviation of vector x. This normalization process can be performed online by feeding the network with the batch of sensory data.

Data Augmentation One challenge for developing a robust skill model with our approach comes from the lack of large-scale data samples in JIGSAWS, where the number of labeled samples is only 40 in total (8 subjects with 5 trial repetitions) for each surgical task. Generally, deep learning might suffer from overfitting if the size of available dataset is limited [33]. To overcome this problem, data augmentation is introduced to prevent overfitting and improve generalization of the deep learning model. This has been seen so far mostly in image recognition, where several methods, such as scaling, cropping, and rotating are used [50,51]. Inspired from the computer vision community, similar augmentation techniques were applied for time series data to enlarge small-sized datasets and increase decoding accuracy [52–54]. In this study, to support the network in learning, we adapted the augmentation strategy and introduced a two-step augmentation process before inputting data into our network. First,

Table 1 Variables of sensory signals from end-effectors of *da Vinci* robot

End-effector category	Description	Variables	Channels
Master tool manipulator (MTM)			
MTM1	Positions (3), rotation matrix (9), velocities (6) of tool tip, gripper angular velocity (1)	$x, y, z, R \in \mathbb{R}^{3\times3}, v_x, v_y, v_z, \omega'_x, \omega'_y, \omega'_z, \alpha$	19×2
MTM2			
Patient-side manipulator (PSM)			
PSM1	Positions (3), rotation matrix (9), velocities (6) of tool tip, gripper angular velocity (1)	$x, y, z, R \in \mathbb{R}^{3\times3}, v_x, v_y, v_z, \omega'_x, \omega'_y, \omega'_z, \alpha$	19×2
PSM2			

These variables are captured as multivariate time series data in each surgical operation trial

Fig. 3 Snapshots of operation tasks during robot-assisted minimally invasive surgical training. The operations are implemented using the *da Vinci* robot and are reported in JIGSAWS [40]: **a** *Suturing*, **b** *Needle-passing*, **c** *Knot-tying*

followed by z-normalization, we viewed and separated the surgical motion data from master (MTMs) and patient-side manipulators (PSMs) as two distinct sample instances, while the class labels for each trial were preserved. This procedure is also appropriate in cases where the MTMs and PSMs are not directly correlated (e.g. position scaling, or other differences in robot control terms). Then, we carried out a label-preserving cropping with a sliding window, where the motion sub-sequences were extracted using crops, i.e., sliding a fixed-size window within the trial. The annotation for each window is identical to the class label of original trial, from which the sub-sequences are extracted. One advantage of this approach is that it leads to larger-scale sets for the robust training and testing of the network. Also, this technique allows us to format time series as equal-length inputs, regardless of the varied lengths of original data. The pseudocode of sliding-window cropping algorithm is shown in Algorithm 1, where X is the input motion data, s is the output crops (i.e., sub-sequences), W is the sliding-window size and L is the step size. After experimenting based on trial-and-error, we chose a window size $W = 60$ and a step size $L = 30$ in this work. Overall, by applying the aforementioned data augmentation process on the original dataset, it resulted in 6290, 6780, and 3542 crops for *Suturing*, *Needle-passing*, and *Knot-tying*, respectively. All of these crops are new data samples for the network. The overall numbers of obtained crops are different since original recording lengths are varied across each trial in JIGSAWS. As a result, we obtained the total sample trials with the size of 6290, 6780, and 3542 for three tasks, respectively, according to the selected setting.

Algorithm 1 Sliding-window Cropping Algorithm

INPUT: raw time series X, *stepSize* L, *windowWidth* W
OUTPUT: sub-sequences $s = \text{SlidingWindow}(X, L, W)$
1: **initialization** $m := 0, n := 0$
2: $s := empty$
3: **while** $m + W \leq length(X)$ **do**
4: $\quad s[n] := X[m : (m + W - 1)]$
5: $\quad m := m + L, n := n + 1$
6: **end while**
7: **return** sub-sequences s

Training & testing

To validate the model classification, we adopt two individual validation schemes in this work: *Leave-one-supertrial-out (LOSO)* and *Hold-out*. The objective of the comparison is to search for the best validation strategy suitable for system development in the case of deep learning. Based on each

cross-validation setting, we train and test a surgical skill model for each surgical task, *Suturing* (SU), *Knot-tying* (KT), and *Needle-passing* (NP).

Leave-one-supertrial-out (LOSO) cross-validation: This technique involves repetitively leaving out a single subset for testing in multiple partitions. Specifically, a supertrial, i, defined as a subset of examples combining the i-th trials from all subjects for a given surgical task [40], is left out for testing, while the union of remaining examples is used for training. This process is repeated in fivefold where each fold consists of each one of the five supertrials. The average of all fivefold performance measures (see "Modeling performance measures" section for definitions) in each test set is reported and gives an aggregated classification result. As a widely used validation strategy, the *LOSO* cross-validation shows its value in evaluating the robustness of a method for skill assessment.

Hold-out: Different from the *LOSO* cross-validation, the *Hold-out* strategy is implemented by conducting a train/test split once, which is normally adopted in deep learning models when large datasets are presented. In this work, one single subset consisting of one of the five trials from each surgeon, for a given surgical task, is left out throughout the training and used as a hold-out for the purpose of testing. Also, to reduce the bias and avoid potential overfitting, we randomly select a trial out of the five repetitions for each subject.

Modeling performance measures

To compare the model performance, classifications are evaluated regarding four common metrics (Eq. 2) [49,55,56]: the average *accuracy*—ratio between the sum of correct predictions and the total number of predictions; *precision*—ratio of correct positive predictions (T_p) and the total positive results predicted by the classifier ($T_\mathrm{p}+F_\mathrm{p}$); *recall*—ratio of positive predictions (T_p) and the total positive results in the ground-truth ($T_\mathrm{p}+F_\mathrm{n}$); and *f1-score*—a weighted harmonic average between *precision* and *recall*.

$$\begin{aligned} precision &= \frac{T_\mathrm{p}}{T_\mathrm{p}+F_\mathrm{p}} \\ recall &= \frac{T_\mathrm{p}}{T_\mathrm{p}+F_\mathrm{n}} \\ f1\text{-}score &= \frac{2*(recall*precision)}{recall+precision} \end{aligned} \tag{2}$$

where T_p and F_p are the numbers of true positives and false positives, T_n and F_n are the numbers of true negatives and false negatives, for a specific class.

In order to assess the computing effort involved in model classification, we measure the running time of skill models to classify all samples in the entire testing set. In the *LOSO* scheme, the running time is measured as the average value from the fivefold cross-validation.

Results

We evaluate the proposed deep learning approach for self-proclaimed skill classification and GRS-based skill classification using the JIGSAWS dataset. The confusion matrices of classification results are obtained from the testing set under the *LOSO* scheme. We compare our results with the state-of-the-art classification approaches in Table 2. It is important to mention that in order to obtain a valid benchmarking analysis, the classifiers investigated in this study are selected among the skill assessment using JIGSAWS motion data and evaluated based on the same *LOSO* validation. Figure 4a shows the results of three-class self-proclaimed skill classification. The proposed deep learning skill model achieved high-accuracy prediction performance. Specifically, our model obtained accuracies of 93.4%, 89.8%, and 84.9% in tasks of *Suturing*, *Needle-passing* and *Knot-tying*, respectively, using a window crop with 2-second duration containing 60 time steps ($W = 60$). In contrast to the per-window assessment, highest accuracies reported in the literature range from 99.9 to 100% via a descriptive model using entropy features based on the entire observation of full operation trial. For the GRS-based skill classification, as shown in Fig. 4b, the proposed approach can achieve higher accuracy than others (92.5%, 95.4%, and 91.3% in *Suturing*, *Needle-passing* and *Knot-tying*). Specifically, the deep learning model outperformed k-nearest neighbors (k-NN), logistic regression (LR), and support vector machine (SVM), with the accuracy improvements ranging from 2.89 to 22.68% in *Suturing*, and 10.94–21.09% in *Knot-tying*.

To study the capability of our proposed approach for online skill decoding, we further evaluate the performance of proposed approach using the input sequences with varying lengths. We repeated our experiment for the self-proclaimed skill classification with different sizes of sliding window: $W1 = 30$, $W2 = 60$ and $W3 = 90$. Modeling performance of window sizes together with the average running time taken for self-proclaimed skill classification is reported in Table 3. The results show that our deep learning skill model can offer advantages over traditional approaches with highly time-efficient skill classification on the per-window basis, without the full observation of surgical motion for each trial (per-trial basis). Also, a higher average accuracy can be found with an increase in sliding-window size. Specifically, the 3-second sliding window containing 90 time steps ($W3 = 90$) can obtain better results compared to 2-second window ($W2 = 60$), with average accuracy improvements of 0.75% in *Suturing*, 0.56% in *Needle-passing* and 2.38% in *Knot-tying*, respectively.

Table 2 Comparison of existing algorithms employed for skill assessment using motion data from JIGSAWS dataset

Author	Algorithm	Labeling approach	Metric extraction	Accuracy			Characteristics
				SU	NP	KT	
Tao et al. [32]	S-HMM	Self-proclaim	Gesture segments	97.4	96.2	94.4	Generative modeling Segment-based Per-trial basis
Forestier et al. [29]	VSM	Self-proclaim	Bag of words features	89.7	96.3	61.1	Descriptive modeling Feature-based Per-trial basis
Zia and Essa [31]	NN	Self-proclaim	Entropy features	100	99.9	100	Descriptive modeling Feature-based Per-trial basis
Fard et al. [24]	*k*-NN	GRS-based	Movement features	89.7	N/A	82.1	Descriptive modeling
	LR			89.9	N/A	82.3	Feature-based Two-class skill only
	SVM			75.4	N/A	75.4	Per-trial basis
Current study	CNN	Self-proclaim	N/A	93.4	89.8	84.9	Deep learning modeling No manual feature
		GRS-based		92.5	95.4	91.3	Per-window basis Online analysis

We benchmark the results in terms of accuracy based on *LOSO* cross-validation. Models conducting classification on the trial level are categorized as *per-trial basis*

Furthermore, in order to characterize the roles of two validation schemes, we repeat the above modeling process using *Hold-out* strategy. Table 3 shows the comparison of self-proclaimed skill classification under *LOSO* cross-validation and *Hold-out* schemes.

Discussion

Recent trends in robot-assisted surgery have promoted a great need for proficient approaches for objective skill assessment [11]. Although several analytical techniques have been developed, efficiently measuring surgical skills from complex surgical data still remains an open problem. In this paper, our primary goal is to introduce and evaluate the applicability of a novel deep learning approach toward online surgical skill assessment. Compared to conventional approaches, our proposed deep learning model reduced dependency on the complex manual feature design or carefully tuned gesture segmentation. Overall, deep learning skill models, with appropriate design choices, yielded competitive performance in both accuracy and time efficiency.

Validity of our deep learning model for objective skill assessment

For results shown in Fig. 4a, b, we note that both *Suturing* and *Needle-passing* are associated with better results than *Knot-tying* in both self-proclaimed skill classification and GRS-based skill classification, indicting that *Knot-tying* is a more difficult task for assessment. For self-proclaimed skill classification, the majority of misclassification errors occurred during the *Knot-tying* task where self-proclaimed $Intermediate$ are misclassified as actual $Novice$. As shown in Fig. 4a, the distribution across $Intermediate$ is pronounced with the probability of 0.34 being misclassified as $Novice$. This could be attributed to the fact that the self-proclaimed skill labels, which are based on hours spent in robot operations, may not accurately reflect the ground-truth knowledge of expertise. As evident, the classification using GRS-based skill labels generally performs better than the results using self-proclaimed skills. Our results indicate that more accurate surgeon skill labels relative to the true surgeon expertise might help to further improve the overall accuracy of skill assessment.

As shown in Table 2, high classification accuracy can be achieved by a few existing methods using generative modeling and descriptive modeling. Specifically, a generative model, sparse HMM (S-HMM), is able to give high predictive accuracy ranging from 94.4 to 97.4%. This result might benefit from a precise description of motion structures and predefined gestures in each task. However, such an approach requires prerequisite segmentation of motion sequences, as well as different complex class-specific models for each skill level [32]. Second, descriptive models sometimes may be

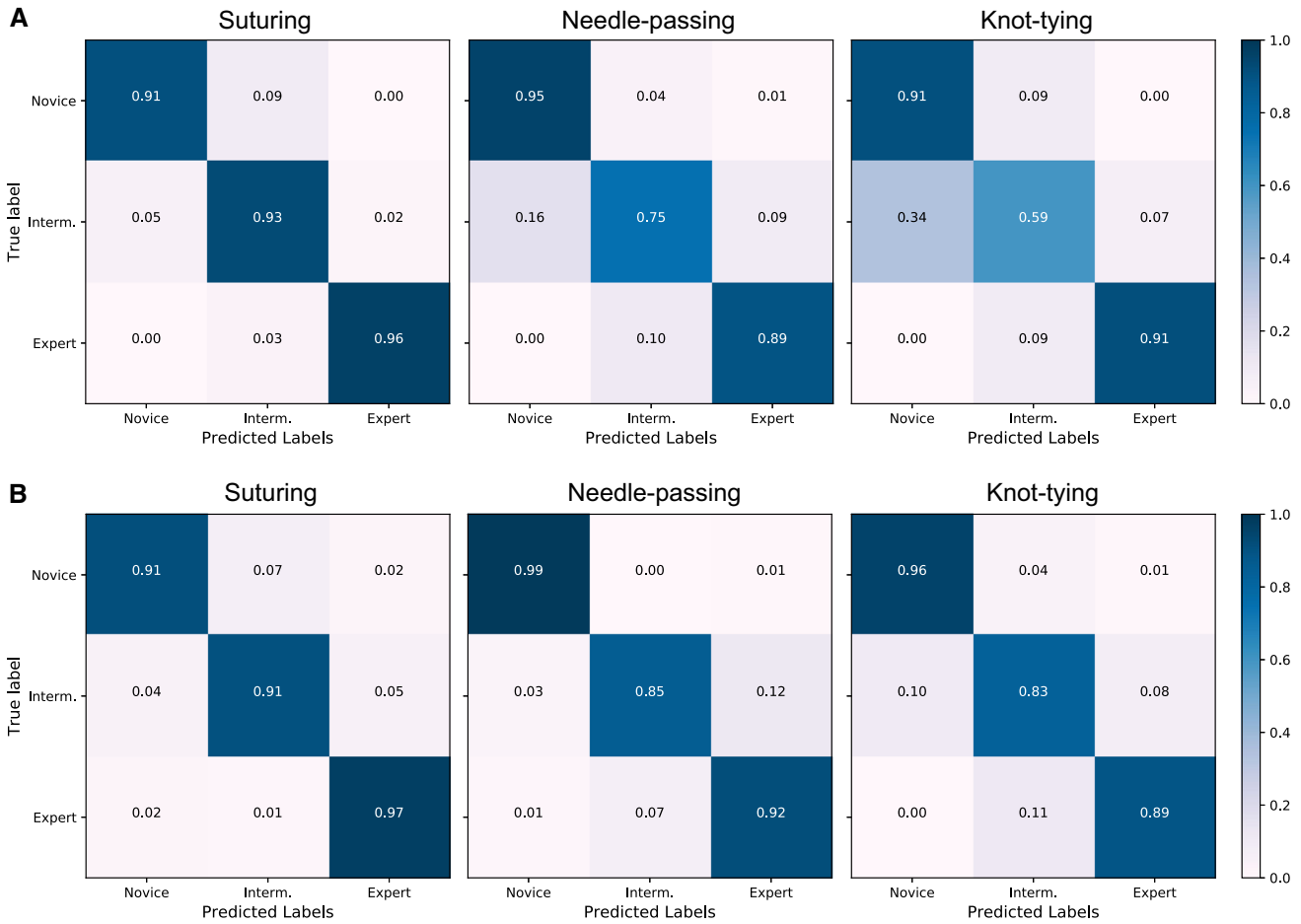

Fig. 4 Confusion matrices of classification results in three surgical training tasks. **a** Self-proclaimed skill classification, **b** GRS-based skill classification. Element value (i, j) and color represent the probability of predicted skill label j, given the self-proclaimed skill label i, where i and $j \in \{1 : \text{“}Novice\text{”}, 2: \text{“}Intermediate\text{”}, 3 : \text{“}Expert\text{”}\}$. The diagonal corresponds to correct predictions

superior to provide highly accurate results, such as the use of novel entropy features. However, the deficiency is that significant domain-specific knowledge and development is required to define the most informative features manually, which directly associate with the final assessment accuracy. This deficiency could also explain why there exists a larger variance in accuracy between other studies (61.1–100%), which are sensitive to the choice of predefined features, as shown in Table 2.

Another attention of our analysis is focused on the optimal sliding windows needed to render an efficient assessment. The duration of time steps in each window should roughly correspond to the minimum time required to decode skills from input signals. Usually, technical skill is assessed at the trial level; however, a quicker and more efficient acquisition may enable immediate feedback to the trainee, possibly improving learning outcomes. Overall, our findings suggest that the per-window-based classification in this work is well-applicable for online settings. Smaller window size can allow for a faster running speed and less delay due to the light-weight computing expense. In contrast, an lager window size implies an increase in delay due to larger network complexity and higher computing effort involved in decoding. Specifically, as shown in Table 3, within the *LOSO* validation scheme, the network can classify the entire testing dataset within 133.88 ms for $W1$ and 172.87 ms running time for $W2$, while it required 214.14 ms running time for $W3$ to classify the samples. However, it is important to mention that given an increase in window sizes, a higher accuracy can be achieved. In particular, there seems to be more gains in the *Knot-tying* analysis, where the highest 2.24% accuracy improvement was obtained from $W2$ to $W3$. This result might be due to the fact that more information of motion dynamics are contained in larger crops, thus allowing for an improved decoding accuracy. We suggest that this trade-off between decoding accuracy and time efficiency could be a factor of interest in online settings of skill assessment.

Table 3 Summary table showing self-proclaimed skill classification performance based on different validation schemes and sliding windows

Task	Validation scheme	Window size	F1-score			Accuracy	Running time (ms)
			Novice	Interm.	Expert		
Suturing	LOSO	$W1$	0.94	0.83	0.95	0.930	**146.45**
		$W2$	0.94	0.83	0.97	0.934	185.40
		$W3$	**0.95**	**0.86**	**0.96**	**0.941**	247.01
	Hold-out	$W1$	0.98	0.92	0.94	0.961	**98.10**
		$W2$	**0.99**	0.94	0.96	0.972	146.40
		$W3$	**0.99**	**0.98**	**0.97**	**0.983**	194.79
Needle-passing	LOSO	$W1$	0.95	0.73	0.88	0.889	**153.36**
		$W2$	0.95	0.75	**0.90**	0.898	194.98
		$W3$	**0.96**	**0.76**	0.89	**0.903**	248.03
	Hold-out	$W1$	0.97	0.80	0.91	0.919	**113.49**
		$W2$	**0.98**	0.81	0.91	0.925	169.72
		$W3$	**0.98**	**0.86**	**0.94**	**0.945**	207.12
Knot-tying	LOSO	$W1$	0.90	0.57	0.90	0.847	**101.83**
		$W2$	0.90	0.62	**0.92**	0.849	138.25
		$W3$	**0.92**	**0.64**	0.91	**0.868**	147.38
	Hold-out	$W1$	0.87	0.42	0.91	0.803	**74.5**
		$W2$	**0.88**	**0.48**	**0.92**	**0.817**	113.55
		$W3$	**0.88**	0.47	0.91	0.816	139.39

Window size is set as $W1 = 30$, $W2 = 60$ and $W3 = 90$. Running time quantifies the computing effort involved in classification. Bold numbers denote best results regarding f1-score, accuracy, and running time

Comparison of validation schemes

We investigated the validity of two different validation schemes for skill modeling. In this case, the differences between both are non-trivial in the deep learning development. Noticeably, *LOSO* cross-validation gives a reliable estimate of system performance. However, the *Hold-out* scheme, which uses a random subset of surgical trials as a hold-out, demonstrates relatively larger variances among results. This result can by explained by the differences among these randomly selected examples in the *Hold-out* validation. Nevertheless, the *Hold-out* shows consistency with the results in *LOSO* scheme across different tasks and window sizes, as shown in Table 3. It is important to note that given a large dataset, the *LOSO* cross-validation might be less efficient for model assessment. In this scenario, the computing load in *LOSO* modeling has been largely increased, which may not be suitable for complex deep architectures. However, the *Hold-out* only needs to run once and is less computationally expensive in modeling.

Limitations

Despite the progress in present work, there still exist some limitations of deep learning models toward a proficient online skill assessment. First, as confirmed by our results, the classification accuracy of supervised deep learning relies heavily on the labeled samples. The primary concern in this study lies with the JIGSAWS dataset and the lack of strict ground-truth labels of skill levels. It is important to mention that there is a lack of consensus in the ground-truth annotation of surgical skills. In the GRS-based labeling, skill labels were annotated based on the predefined cutoff threshold of GRS scores, however, no commonly accepted cutoff exists. For future work, a refined labeling approach with stronger ground-truth knowledge of surgeon expertise may further improve the overall skill assessment [57,58]. Second, we will search for a detailed optimization of our deep architecture, parameter settings, and augmentation strategies to better handle motion time series data and improve the online performance further. In addition, the interpretability of automatically learned representations is currently limited due to the black-box nature of deep learning models. It would be interesting to investigate a visualization of deep hierarchical representations to understand hidden skill patterns, so as to better justify the decision taken by a deep learning classifier.

Conclusion

The primary contributions of this study are: (1) a novel data-driven deep architecture for an active classification of

surgical skill via end-to-end learnings, (2) an insight in accuracy and time efficiency improvements for online skill assessment, and (3) application of data augmentation and exploration of validation schemes feasible for deep skill modeling. Taking advantage of recent technique advances, our approach has several desirable proprieties and is extremely valuable for online skill assessment. First, a key benefit is an end-to-end skill decoding, learning abstract representations of surgery motion data with automatic recognitions of skill levels. Without a priori dependency on engineered features or segmentation, the proposed model achieved comparable results to previously reported methods. It yielded highly competitive time efficiency given relatively small crops (1–3 second window with 30–90 time steps), which were computationally feasible for online assessment and immediate feedback in training. Furthermore, we demonstrated that an improvement of modeling performance could be achieved by the optimization of design choices. An appropriate window size could provide better results in *Knot-tying* with a 2.24% accuracy increase. Also, the development of deep skill models might benefit from the *Hold-out* strategy, which requires less computing effort than the *LOSO* cross-validation, especially in the case where large datasets are involved.

Overall, the ability to automatically learn abstract representations from raw sensory data with high predictive accuracy and fast processing speed, makes our approach well-suited for online objective skill assessment. The proposed deep model can be easily integrated into the pipeline of robot-assisted surgical systems and could allow for immediate feedback in personalized surgical training.

Acknowledgements This work is supported by National Science Foundation (NSF#1464432).

Compliance with ethical standards

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

Ethical approval For this type of study formal consent is not required.

Informed consent This articles does not contain patient data.

References

1. Roberts KE, Bell RL, Duffy AJ (2006) Evolution of surgical skills training. World J Gastroenterol WJG 12(20):3219
2. Reznick RK, MacRae H (2006) Teaching surgical skills changes in the wind. N Engl J Med 355(25):2664–2669
3. Aggarwal R, Mytton OT, Derbrew M, Hananel D, Heydenburg M, Issenberg B, MacAulay C, Mancini ME, Morimoto T, Soper N, Ziv A, Reznick R (2010) Training and simulation for patient safety. BMJ Qual Saf 19(Suppl 2):i34–i43
4. Birkmeyer JD, Finks JF, O'reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ, (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
5. Darzi A, Mackay S (2001) Assessment of surgical competence. BMJ Qual Saf 10(suppl 2):ii64–ii69
6. Bridgewater B, Grayson AD, Jackson M, Brooks N, Grotte GJ, Keenan DJ, Millner R, Fabri BM, Mark J (2003) Surgeon specific mortality in adult cardiac surgery: comparison between crude and risk stratified data. Br Med J 327(7405):13–17
7. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187(1):247–252
8. Aghazadeh MA, Jayaratna IS, Hung AJ, Pan MM, Desai MM, Gill IS, Goh AC (2015) External validation of global evaluative assessment of robotic skills (gears). Surg Endosc 29(11):3261–3266
9. Niitsu H, Hirabayashi N, Yoshimitsu M, Mimura T, Taomoto J, Sugiyama Y, Murakami S, Saeki S, Mukaida H, Takiyama W (2013) 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 43(3):271–275
10. Reiley CE, Lin HC, Yuh DD, Hager GD (2011) Review of methods for objective surgical skill evaluation. Surg Endosc 25(2):356–366
11. Vedula SS, Ishii M, Hager GD (2017) Objective assessment of surgical technical skill and competency in the operating room. Ann Rev Biomed Eng 19:301–325
12. Moustris GP, Hiridis SC, Deliparaschos KM, Konstantinidis KM (2011) Evolution of autonomous and semi-autonomous robotic surgical systems: a review of the literature. Int J Med Rob Comput Assist Surg 7(4):375–392
13. Cheng C, Sa-Ngasoongsong A, Beyca O, Le T, Yang H, Kong Z, Bukkapatnam ST (2015) Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. IIE Trans 47(10):1053–1071
14. Klonowski W (2009) Everything you wanted to ask about eeg but were afraid to get the right answer. Nonlinear Biomed Phys 3(1):2
15. Reiley CE, Hager GD (2009) Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 435–442
16. Kassahun Y, Yu B, Tibebu AT, Stoyanov D, Giannarou S, Metzen JH, Vander Poorten E (2016) Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. Int J Comput Assist Radiol Surg 11(4):553–568
17. Judkins TN, Oleynikov D, Stergiou N (2009) Objective evaluation of expert and novice performance during robotic surgical training tasks. Surg Endosc 23(3):590
18. Liang K, Xing Y, Li J, Wang S, Li A, Li J (2018) Motion control skill assessment based on kinematic analysis of robotic end-effector movements. Int J Med Rob Comput Assist Surg 14(1):e1845-n/a. https://doi.org/10.1002/rcs.1845
19. Trejos AL, Patel RV, Malthaner RA, Schlachta CM (2014) Development of force-based metrics for skills assessment in minimally invasive surgery. Surg Endosc 28(7):2106–2119
20. Poursartip B, LeBel M-E, Patel R, Naish M, Trejos AL (2017) Analysis of energy-based metrics for laparoscopic skills assessment. IEEE Trans Biomed Eng 65(7):1532–1542
21. Ershad M, Koesters Z, Rege R, Majewicz A (2016) Meaningful assessment of surgical expertise: semantic labeling with data and crowds. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 508–515

22. Sharon Y, Lendvay TS, Nisky I (2017) Instrument orientation-based metrics for surgical skill evaluation in robot-assisted and open needle driving. arXiv preprint arXiv:1709.09452
23. Shackelford S, Bowyer M (2017) Modern metrics for evaluating surgical technical skills. Curr Surg Rep 5(10):24
24. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, Klein MD (2018) Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Rob Comput Assist Surg 14(1):e1850. https://doi.org/10.1002/rcs.1850
25. Stefanidis D, Scott DJ, Korndorffer JR Jr (2009) Do metrics matter? Time versus motion tracking for performance assessment of proficiency-based laparoscopic skills training. Simul Healthc 4(2):104–108
26. Chmarra MK, Klein S, de Winter JC, Jansen F-W, Dankelman J (2010) Objective classification of residents based on their psychomotor laparoscopic skills. Surg Endosc 24(5):1031–1039
27. Vedula SS, Malpani A, Ahmidi N, Khudanpur S, Hager G, Chen CCG (2016) Task-level vs. segment-level quantitative metrics for surgical skill assessment. J Surg Educ 73(3):482–489
28. Poursartip B, LeBel M-E, McCracken LC, Escoto A, Patel RV, Naish MD, Trejos AL (2017) Energy-based metrics for arthroscopic skills assessment. Sensors 17(8):1808
29. Forestier G, Petitjean F, Senin P, Despinoy F, Jannin P (2017) Discovering discriminative and interpretable patterns for surgical motion analysis. In: Conference on artificial intelligence in medicine in Europe. Springer, pp 136–145
30. Brown JD, OBrien CE, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ, (2017) Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. IEEE Trans Biomed Eng 64(9):2263–2275
31. Zia A, Essa I (2018) Automated surgical skill assessment in RMIS training. Int J Comput Assist Radiol Surg. https://doi.org/10.1007/s11548-018-1735-5
32. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R (2012) Sparse hidden markov models for surgical gesture classification and skill evaluation. In: IPCAI. Springer, pp 167–177
33. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
34. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117
35. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484–489
36. Graves A, Mohamed A-R, Hinton G (2013) Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6645–6649
37. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118
38. Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY (2017) Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836
39. DiPietro R, Lea C, Malpani A, Ahmidi N, Vedula SS, Lee GI, Lee MR, Hager GD (2016) Recognizing surgical activities with recurrent neural networks. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 551–558
40. Gao Y, Vedula SS, Reiley CE, Ahmidi N, Varadarajan B, Lin HC, Tao L, Zappella L, Béjar B, Yuh DD, Chen CCG, Vidal R, Khudanpur S, Hager GD (2014) JHU-ISI gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling. In: MICCAI workshop: M2CAI, vol 3, p 3
41. Längkvist M, Karlsson L, Loutfi A (2014) A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognit Lett 42:11–24
42. Gamboa JCB (2017) Deep learning for time-series analysis. arXiv preprint arXiv:1701.01887
43. Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
44. Chollet F (2015) Keras. https://keras.io. Accessed 12 Dec 2017
45. Li M, Zhang T, Chen Y, Smola AJ (2014) Efficient mini-batch training for stochastic optimization. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 661–670
46. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980
47. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
48. Wu H, Gu X (2015) Max-pooling dropout for regularization of convolutional neural networks. In: International conference on neural information processing. Springer, pp 46–54
49. Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, Hager GD (2017) A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Trans Biomed Eng 64(9):2025–2041
50. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classication with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems. Curran Associates Inc., NIPS, vol 1, pp 1097–1105
51. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
52. Cui Z, Chen W, Chen Y (2016) Multi-scale convolutional neural networks for time series classification. arXiv preprint arXiv:1603.06995
53. Le Guennec A, Malinowski S, Tavenard R (2016) Data augmentation for time series classification using convolutional neural networks. In: ECML/PKDD workshop on advanced analytics and learning on temporal data
54. Um TT, Pfister FM, Pichler D, Endo S, Lang M, Hirche S, Fietzek U, Kulić D (2017) Data augmentation of wearable sensor data for Parkinsons disease monitoring using convolutional neural networks. In: Proceedings of the 19th ACM international conference on multimodal interaction. ACM, pp 216–220
55. Sammut C, Webb GI (2011) Encycl Mach Learn. Springer, Berlin
56. Kumar R, Jog A, Malpani A, Vagvolgyi B, Yuh D, Nguyen H, Hager G, Chen CCG (2012) Assessing system operation skills in robotic surgery trainees. Int J Med Rob Comput Assist Surg 8(1):118–124
57. Sun C, Shrivastava A, Singh S, Gupta A (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: 2017 IEEE international conference on computer vision (ICCV). IEEE, pp 843–852
58. Dockter RL, Lendvay TS, Sweet RM, Kowalewski TM (2017) The minimally acceptable classification criterion for surgical skill: intent vectors and separability of raw motion data. Int J Comput Assist Radiol Surg 12(7):1151–1159