Video-based assessment of intraoperative surgical skill

Sanchit Hira1 · Digvijay Singh2 · Tae Soo Kim2,3 · Shobhit Gupta4 · Gregory Hager1,2,3 · Shameema Sikder1,3,5 · S. Swaroop Vedula3

Abstract

Purpose Surgeons’ skill in the operating room is a major determinant of patient outcomes. Assessment of surgeons’ skill is necessary to improve patient outcomes and quality of care through surgical training and coaching. Methods for video-based assessment of surgical skill can provide objective and efficient tools for surgeons. Our work introduces a new method based on attention mechanisms and provides a comprehensive comparative analysis of state-of-the-art methods for video-based assessment of surgical skill in the operating room.

Methods Using a dataset of 99 videos of capsulorhexis, a critical step in cataract surgery, we evaluated image feature-based methods and two deep learning methods to assess skill using RGB videos. In the first method, we predict instrument tips as keypoints and predict surgical skill using temporal convolutional neural networks. In the second method, we propose a frame-wise encoder (2D convolutional neural network) followed by a temporal model (recurrent neural network), both of which are augmented by visual attention mechanisms. We computed the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and predictive values through fivefold cross-validation.

Results To classify a binary skill label (expert vs. novice), the range of AUC estimates was 0.49 (95% confidence interval; CI = 0.37 to 0.60) to 0.76 (95% CI = 0.66 to 0.85) for image feature-based methods. The sensitivity and specificity were consistently high for none of the methods. For the deep learning methods, the AUC was 0.79 (95% CI = 0.70 to 0.88) using keypoints alone, 0.78 (95% CI = 0.69 to 0.88) and 0.75 (95% CI = 0.65 to 0.85) with and without attention mechanisms, respectively.

Conclusion Deep learning methods are necessary for video-based assessment of surgical skill in the operating room. Attention mechanisms improved discrimination ability of the network. Our findings should be evaluated for external validity in other datasets.

Keywords Video-based assessment · Surgical skill · Deep learning · Cataract surgery

Introduction

Surgeons’ skill in the operating room affects patient outcomes [5]. Interventions to optimize surgeons’ skill can improve quality of patient care. Assessment of skill is a cornerstone for any intervention to improve it. Surgical skill assessment is essential for other purposes throughout surgeons’ career including in-training evaluations, surgical coaching, certification, re-certification, and credentialing of surgeons, as well as end of career decisions [18].

Traditionally, surgical skill assessment was based upon direct observation of surgeons in the operating room [29].
which was subjective and unreliable. Unlike direct observation, video-based assessment (VBA) enables asynchronous objective evaluation of surgeons’ skill, provides surgeons with formative assessments, and enables surgical coaching among other applications [24,28].

Currently, VBA of surgical skill in the operating room is obtained either from ratings by peer surgeons or crowd raters (i.e., crowd-sourcing) [24]. Manual VBA of surgical skill is subjective and unreliable. Crowd-sourced skill ratings correlate with expert ratings [16,19,23], but some studies show low predictive value for the ratings [1,8]. Consequently, despite the efficiency with which skill ratings can be obtained with crowdsourcing, its role in routine assessments of surgical skill is not established. Furthermore, VBA by peers relies upon access to experienced raters and thus, it is inefficient for routine use.

Surgical data science methods can provide objective and accurate VBA of surgical skill [18]. Previously, we developed and evaluated a temporal convolutional neural network (TCN) to assess skill using manually annotated instrument tips in video images [13]. Our objective in this study is to develop and validate methods for assessment of surgical skill directly from videos of the surgical field. Using a dataset of 99 videos, we evaluated algorithms to assess surgical skill with a uniform cross-validation setup and evaluation metrics. For VBA of skill in the operating room, to our knowledge, this dataset is the largest in the literature. The contributions of our work include:

1. A comprehensive evaluation of feature-based methods for VBA of intraoperative surgical skill;
2. Methods for surgical skill assessment directly from RGB videos;
3. A novel architecture for a deep learning method augmented with attention to analyze surgical videos, toward explaining the predictions.

Methods

Feature-based approaches

We evaluate five approaches to obtain features from videos, which we then analyze with different classifiers, including support vector machines (SVMs) with linear and radial basis frequency kernels, logistic regression, and a multilayer perceptron (MLP).

Detectors

Following [15], we use spatiotemporal interest points (STIPs) to identify regions in images that are key to developing useful feature representations.

Descriptors

We compute three different descriptors for each STIP: histogram of oriented gradients (HoG) to encode local spatial information in image patches [7], histogram of optical flow (HoF) to describe localized flow in videos [7], and motion boundary histograms (MBHs) to encode localized temporal information in X and Y components of the differential flow [35].

Features

Bag of Words (BoW) We cluster descriptors of STIPs in a video using a k-means algorithm to obtain a visual feature vocabulary, and use TF-IDF to compute the BoW feature [25].

Augmented Bag of Words (Aug. BoW) To address lack of temporal information in BoW, we create a temporal vocabulary by classifying the time between STIPs into N bins. We concatenate the visual and temporal vocabularies and compute n-grams. The n-grams are thus a sequence of events and sum of their times. This is a useful representation for surgical movements that take relatively longer periods of time and continue over a number of frames. We use TF-IDF to represent n-grams [4].

Discrete Fourier Transform (DFT) / Discrete Cosine Transform (DCT) DFT and DCT extract motion information, i.e., the frequencies of the different surgical action categories [33]. A transformation matrix or motion class matrix (MCM) is created from the learned clusters of the visual vocabulary using K-means, so that if a video has N frames, a KxN MCM is created that contains for each row i = 1,2,…, K, the counts of how many interest points in the nth frame belong to the cluster K. The DFT and DCT of this matrix are then calculated, where each entry in the calculated KxN matrix represents the nth frequency coefficient of the k-th cluster. This matrix measures the repetitiveness of the different surgical actions in the videos. Since higher frequencies are a result of noisy or abrupt movements, the lowest D frequencies are used, and we get a KxD matrix. This matrix is flattened to get a K*D feature vector.

Sequential Motion Textures (SMT) In SMT, spatiotemporal information is encoded in the form of gray level co-occurrence matrices (GLCMs) that are derived from frame kernel matrices of the MCM [26]. They represent the affinity of the elements of a matrix with respect to all the elements. The frame kernel matrices are obtained by splitting the MCM into time windows of specified width (W), which allows temporal information to be considered over a fixed time interval according to the length of the videos. This is followed by applying a radial basis frequency kernel and shifting the domain to grayscale. Texture patterns are then extracted from these matrices using GLCMs for differ-
ent gray levels Ng, which encode both spatial relationships and motion dynamics in the surgical videos. After calculating the averaged and normalized GLCMs for the windows, 20 standard texture features are selected using sequential forward feature selection; this results in Wx20 features for each video [20].

**Approximate Entropy (ApEn) / Cross Approximate Entropy (XApEn)** ApEn and XApEn are methods that construct entropy-based features that measure the amount of entropy in a given time-series input [34]. These features are expected to recognize predictable and regular patterns in time series which in turn leads to better skill assessment from video-based descriptors and accelerometer data [34]. For ApEn, each time series is split into embedding factors (time windows) according to a time delay, and then for each embedding factor the frequency of repeatable patterns is calculated by summing up the Heaviside functions of the embedding factor the frequency of repeatable patterns is according to a time delay, and then for each embedding factor the frequency of repeatable patterns is calculated by summing up the Heaviside functions of the L-Infinity norm distance MCs [34]. XApEn calculation is similar to ApEn, except that the L-Infinity norm distance is calculated between embedding factors for all possible cluster pairs.

**Tool detection approach for skill assessment**

Prior work shows that tool motion data is useful for surgical skill assessment [13], and tooltip annotations in video images can be obtained through crowdsourcing. To avoid the need for tooltip annotation, we learn to detect tooltips as keypoints of objects in video and analyze the predicted keypoints using a TCN (KP-TCN).

**Detection of surgical instrument tips**

We use a TCN described in [13]. We include an additional stage to infer tooltip locations directly from images. We model surgical tool tip locations as keypoints of objects. Let \( X = \{x_1, x_2, \ldots, x_N\} \) be a video with \( N \) images. We assume there exists annotations regarding tooltip locations \( y_n = \{p^n_1, \ldots, p^n_K\} \) for each frame \( x_n \) where \( p^n_k \in \mathbb{R}^2 \) is the pixel location of the \( k \)-th keypoint in frame \( n \). We learn a keypoint detector \( \mathcal{F} \) that minimizes the following objective:

\[
[q^n_1, \ldots, q^n_K] = \mathcal{F}(x_n)
\]

\[
L_{keypoints} = \sum_{n=1}^{N} \sum_{k=1}^{K} d(q^n_k, p^n_k)
\]

where \( d \) is a distance between the keypoints and \( q^n_k \in \mathbb{R}^2 \) is the keypoint prediction given \( x_n \). In this work, we use a convolutional neural network-based keypoint detector for \( \mathcal{F} \) and train it in an end-to-end manner.

**TCNs**

Using \( \mathcal{F} \), we encode a video \( X = \{x_1, x_2, \ldots, x_N\} \) as a time series of keypoint predictions \( Z \) such that:

\[
Z = \{z_1, z_2, \ldots, z_N\} \in \mathbb{R}^{N \times 2K}
\]

where \( z_n = \mathcal{F}(x_n) \in \mathbb{R}^{2K} \) is the predicted pixel locations of \( K \) keypoints in frame \( x_n \).

We calculate tooltip velocities as the difference in predicted keypoint locations between successive frames as in [13].

The final input to the TCN is represented as \( Z \) such that:

\[
Z = \{\delta z_1, \delta z_2, \ldots, \delta z_N\}
\]

where \( \delta z_n = z_{n+1} - z_n \) is the tooltip velocity at frame \( n \).

We follow the TCN design proposed in [13]. For the \( l^{th} \) layer of the TCN with \( F_l \) 1-D convolutional filters, the output activation can be written as

\[
h^l = \sigma(W_l \ast h^{l-1})
\]

where \( h^l \) is the output of the \( l\)-th layer, \( W_l \) are the convolutional filter weights of the \( l^{th} \) layer, \( \sigma(\cdot) \) is a nonlinearity (ReLU followed by batch-normalization operation) and \( \ast \) represents convolution. We train the TCN in an end-to-end manner using standard cross-entropy loss.

**Dual attention network (ATT)**

The TCN omits important contextual information such as visual changes in anatomy and instrument/anatomy interactions. To address this limitation, we present an attention-based method for VBA directly from RGB videos. Past studies have shown that attention mechanisms generate better context vectors that localize to relevant parts of the input for the task [3]. We hypothesize that including attention in our model would enable us to localize time-frames and locations in the video which might be useful for classification. Our architecture includes a feature extractor and an LSTM module, both equipped with separate but dependent spatial and temporal attention mechanisms as explained below.

**Video representation**

First, a ResNet encoder [11] is used to extract features from a given sequence of RGB frames. Then, a LSTM cell [12] operates on those features to generate the final classification. We augment both the ResNet encoder and the LSTM Cell with attention [30], [31].

For the spatial domain, the ResNet returns D feature maps of size \( H \times W \) for each image in a video. The output of the
feature extractor is represented by

\[ V = \{ F_1, F_2, \ldots, F_N \}, V \in \mathbb{R}^{N \times D \times H \times W} \] (5)

where \( F_n \in \mathbb{R}^{D \times H \times W} \) is a spatial feature and \( N \) is the number of frames in video. The frame features in turn consist of a \( D \)-dimensional positional activation map, represented as follows:

\[ F_n = \{ a_{n1}, a_{n2}, \ldots, a_{nL} \}, \quad a_{ni} \in \mathbb{R}^D \] (6)

where \( D \) is the dimension of the feature vector for each position in the image, and \( L = H \times W \) is the number of positions in the image.

**Spatial and temporal attention**

The frame feature vectors \( F_n \) are passed sequentially through the LSTM cell, and then both the hidden state of the LSTM cell and the frame feature vectors are used to compute the attention weights as follows:

\[ e_{ni} = f_{att}(F_n, h_{n-1}) = f_{att}(\{ a_{n1}, a_{n2}, \ldots, a_{nL} \}, h_{n-1}) \]

\[ \alpha_{ni} = \frac{\exp(e_{ni})}{\sum_{j=1}^{L} \exp(e_{nj})} \] (7)

where \( \alpha_{ni} \) are the positional attention weights, and \( f_{att} \) is the spatial attention module, as described below.

\[ att_f = F_n \times W_f \]
\[ att_l = h_{n-1} \times W_h \]

\[ f_{att}(F_n, h_{n-1}) = ReLU(\text{att}_f + \text{att}_l) \times W_c \] (8)

where \( W_f, W_h \) and \( W_c \) represent linear layers. Once the attention weights \( \alpha_{ni} \) are computed, the final context vector \( z_n \) can be computed as follows:

\[ z_n = \sum_{i=1}^{L} \alpha_{ni} \cdot a_{ni} \] (9)

The attention weighted frame encoding vectors are then used to compute the attended feature vector from the LSTM outputs as shown in eq. 10.

\[ Z = \{ z_1, z_2, \ldots, z_N \} \]
\[ M = h_N \times \tanh(Z) \]
\[ \beta = \text{softmax}(M) \]
\[ r = H \beta \] (10)

where \( z_i \) is the spatial attention weighted frame encoding for frame \( i \), \( M \) are the soft alignment scores, \( \beta \) are the temporal attention weights, \( H \) is the combined LSTM output for all frames, and \( r \) is the final temporal attention weighted encoding for the entire video.

Finally, the attended feature encoding is input to a linear classification layer to generate the final classification. We use cross-entropy loss and train the models for 1000 epochs with a step decay on plateau using stochastic gradient descent with an initial learning rate of \( 1e^{-2} \).

**Dataset**

We used a dataset of 99 videos of capsulorhexis (a critical step in cataract surgery), which is identical to that used in [13]. The Johns Hopkins Medical Institutions Institutional Review Board approved this study. A faculty surgeon operated in 28 instances and a trainee surgeon operated in 71 instances. We processed videos captured from the operating microscope to a resolution of 640*480 at 59 frames per second.

One expert surgeon assigned a rating for skill using the International Council of Ophthalmology’s Ophthalmology Surgical Competency Assessment Rubric for phacoemulsification (ICO-OSCAR:phaco) [9]. ICO-OSCAR:phaco includes two items to assess skill for capsulorhexis corresponding to surgical two goals in this step—commencement of flap & follow through (CF), and capsulorhexis formation and completion (RF). Each item is rated on a scale of 2–5.

**Experiments**

We performed two experiments—predicting a binary skill class (expert/novice) and predicting a score on CF and RF. For the binary skill class label, videos assigned a score of 5 on at least one item for capsulorhexis in ICO-OSCAR:phaco.
and at least a score of 4 on the other item were labeled expert; otherwise, the videos were labeled novice. To obtain a 3-class label for CF and RF, which are scored as 2, 3, 4, or 5, we collapsed scores of 2 and 3 into one class for our analysis. We used the same fivefold cross-validation folds setup for both the experiments (Table 1). Holding each fold out at a time for testing, we iteratively used one of the remaining four folds for validation (hyperparameter selection) and trained on the rest of the three folds. We used the model configuration and hyperparameters for which accuracy on the validation fold was highest.

Implementation

Feature-based approaches

To compute STIPs, we use windows of 6 s with a 2 s overlap on each side. The frame rate for input videos is 59 fps, resulting in 520 frames for each window. We use a 3 × 3 × 3 Gaussian kernel with spatial variance of 4 and temporal variance of 8 for the convolution. We use a Gaussian kernel with variance 1 to smooth the response function. We used the top 1000 STIPs.

For HoGs, we extracted a 9 × 9 × 9 patch centered on each STIP, split it into 9 windows, and computed 8 gradient orientations resulting in a 72-dimensional feature vector. For HoFs, we extracted a 17 × 17 patch centered on each STIP. We computed optical flow within the patch, and split it into 9 windows to compute 9 gradient orientations to obtain an 81-dimensional vector. For MBHs, we used 17 × 17 patches separately in the X and Y components of the optical flow that we split into 25 windows and compute 8 gradient orientations (HoGs). This gives a 400-dimensional vector (after concatenating the vector for X and Y components). We concatenated all three descriptors resulting in a 513-dimensional feature for each video.

For k-means in Aug. BoW, we evaluated \( K = 25, 50, \) and 100 and using Euclidean and Mahalanobis distance. We computed n-grams using interspersed, cumulative, or pyramid encoding, where \( n = 3 \) and 5. In interspersed encoding, the temporal information is the time between events (or cluster labels), and this is encoded in sequence of the cluster labels occurring in the videos. This is useful for short surgical movements that take small amounts of time and are independent in the frames of the videos. Cumulative encoding encodes temporal information over a (user specified) sequence of events, and sums this up. Pyramid encoding creates l-grams for all of the specified n-grams from 1 to \( n \). This can be interpreted as breaking down surgical movements and representing them at different levels. We used 5 bins to create the temporal vocabulary, following previous recommendation [4,32].

For DFT and DCT, the frequencies for each cluster time series in the MCM are calculated, and the 50 lowest frequencies are selected from each to ignore noisy frequencies. This gives a \( K × 50 \) vector for each of DFT and DCT, resulting in a \( 2 × K × 50 \)-dimensional vector. The top 30 features are then selected using the SFFS feature selection method, yielding a final 30-dimensional frequency feature vector. SFFS works by evaluating the effectiveness of each individual feature using a classifier, and selecting the top \( n \) (specified by user) features.

For SMT, we analyzed multiple frame kernel matrix variances including 1e-8, 1e-7, 1e-6, 1e-5, 1e-4, and 1e-3. We evaluated multiple window sizes (W) of 2, 4, 8, and 16. We also analyzed multiple gray levels (Ng) of 8, 16, 32, 64, 128, and 256.

For entropy-based methods, we combined features from ApEn and XApEn. The dimensionality of features for ApEn is \( r × K \), and for XApEn is \( r × K × (K − 1)/2 \), where \( K \) is the number of clusters, and ApEn and XApEn features are concatenated to produce a single feature vector. We have \( r \) fixed at 1 (according to [32]), and \( m \) taking values of 1 and 2. \( r_{coeff} \) takes values 0.1, 0.15, 0.2, 0.25, where \( r = r_{coeff} × \sigma \) where \( \sigma \) is the standard deviation of the time series.

We implemented logistic regression with L2 penalty. We chose hyperparameters for the classifiers, including regularization for SVMs and logistic regression, and the number of hidden layers in the MLP, using the validation fold.

Deep learning approaches

All deep learning approaches are implemented using the PyTorch [22] framework and trained using K80 GPUs.

For tool detection, we trained the HR-Net [27] using tool tip annotations of [13]. We use standard image augmentations during training which includes horizontal flip, shift-scale-rotate, brightness, and contrast adjustments. We follow the optimization proposed in [27] and use weighted focal loss [17] with a positive weight of 3 and 1-cycle learning rate schedule with an initial learning rate of 1e-2. We used the Adam optimizer for training.

For TCNs, we use the Adam optimizer with a learning rate of 1e-3 and models train for 50 epochs using standard cross-entropy loss [14]. To train the dual attention network, we sample video snippets of 64 frames, sampled with a stride of 4 frames from each video. The image frames are normalized, and flip and rotate data augmentations are applied. The models are trained for 100 epochs with Adam optimizer with an initial learning rate of 1e-2 that is lowered by a factor of 10 as validation loss plateaus. We used standard cross-entropy loss.

At test time using dual attention networks, a video is represented as clips of 64 frames sampled 4 frames apart to be consistent with training. The clips overlap by 32 frames. Predictions are made individually on all the clips and averaged to produce the final output.
We also evaluated a network trained with without attention (No ATT), using the same implementation described above for the dual attention network.

**Statistical analyses**

To evaluate algorithm performance in each experiment, we computed the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive and negative predictive values (PPV, NPV), and accuracy (both micro- and macro-accuracy [13]). We used bootstrap to compute 95% confidence intervals (CI) for AUC and computed the Wilson interval for the remaining measures [2]. We used the method described in [10] to estimate AUC and bootstrap to compute 95% CI for algorithms predicting a score on each item for capsulorhexis in ICO-OSCAR:phaco.

**Results**

**Feature-based approaches** To predict an expert/novice label, findings using a linear SVM classifier are shown in Figs. 1 and 2 and Table 2. Findings for other classifiers are shown in Table 2 in the Supplementary Material.

None of the features enabled consistently high measures of performance for any of the classifiers. The estimates of AUC were 0.75 or greater for some features and classifiers, suggesting no meaningful difference from deep learning models. However, classifier performance was not uniform for positive (sensitivity) and negative labeled instances (specificity). The utility of the classifiers is affected by low values of sensitivity or specificity.

Among the features we computed, the AUC across classifiers was consistently low (0.49 to 0.54) for SMT, and high for Aug. BoW (0.64 to 0.75) and ApEN (0.67 to 0.73). Furthermore, none of the features yielded a classifier with uniform sensitivity and specificity. These findings suggest that there is limited information in the features we evaluated for surgical skill assessment using intraoperative videos.

To predict a 3-class label for CF and RF using a linear SVM, estimates of AUC for all the feature-based methods were consistent with the null value, although estimates of micro- and macro-accuracy, sensitivity, and specificity indicate that algorithm performance may have been affected by...
the class imbalance in our dataset (Tables 3 and 4 in Supplementary Material).

**Deep learning methods** To predict an expert / novice label, the attention-based network had high AUC, in addition to higher sensitivity and specificity than the other two deep learning methods (Figs. 3 and 4). KP had similar AUC as the attention-based network, but it had lower sensitivity and specificity.

For predicting a 3-class label for CF and RF, estimates of AUC were lower than that for expert / novice label prediction (Table 2). In addition, these methods had higher sensitivity and lower specificity for labels indicating better skill, and lower sensitivity and higher specificity for labels indicating poor skill (Table 3). This observation, together with estimates of micro- and macro-accuracy, suggests that performance of deep learning methods was affected by class imbalance in CF and RF in our dataset.

**Discussion**

Our work is a comprehensive evaluation of state-of-the-art methods for VBA of surgical skill in the operating room. Our findings show that deep learning methods perform better than most feature-based methods in terms of AUC. Furthermore, a network using attention mechanisms had the most desirable performance measures for VBA of skill directly from RGB videos of the surgical field. Even when compared with a network trained using precise manual annotations of instrument tips [13], we observed higher sensitivity (0.843 vs. 0.824) and specificity (0.75 vs. 0.71) with the attention-based network. These findings indicate that it is useful to analyze the
**Table 3** Estimates of performance measures for scores on individual items in ICO-OSCAR:phaco for capsulorhexis (95% confidence intervals in parentheses)

| Item | Sensitivity | Specificity | PPV | NPV |
|------|-------------|-------------|-----|-----|
| **KP: predicted keypoints analyzed with a TCN** | | | | |
| CF = 2/3 | 0.00 (0.00–0.22) | 0.99 (0.94–1.00) | 0.00 (0.00–0.95) | 0.86 (0.77–0.91) |
| CF = 4 | 0.64 (0.48–0.77) | 0.72 (0.59–0.81) | 0.60 (0.44–0.73) | 0.75 (0.63–0.85) |
| CF = 5 | 0.85 (0.72–0.92) | 0.68 (0.55–0.79) | 0.70 (0.57–0.80) | 0.84 (0.70–0.92) |
| RF = 2/3 | 0.35 (0.19–0.55) | 0.93 (0.86–0.97) | 0.62 (0.36–0.82) | 0.83 (0.73–0.89) |
| RF = 4 | 0.45 (0.30–0.62) | 0.80 (0.69–0.88) | 0.54 (0.36–0.70) | 0.75 (0.63–0.83) |
| RF = 5 | 0.84 (0.70 to 0.92) | 0.61 (0.48 to 0.72) | 0.62 (0.49 to 0.73) | 0.83 (0.69–0.91) |
| **ATT: neural network with attention mechanisms** | | | | |
| CF = 2/3 | 0.07 (0.00–0.31) | 0.99 (0.94–1.00) | 0.50 (0.03–0.97) | 0.87 (0.78–0.92) |
| CF = 4 | 0.62 (0.46–0.75) | 0.73 (0.61–0.83) | 0.60 (0.45–0.74) | 0.75 (0.62–0.84) |
| CF = 5 | 0.83 (0.69 to 0.91) | 0.64 (0.51 to 0.76) | 0.67 (0.54 to 0.78) | 0.81 (0.67–0.90) |
| RF = 2/3 | 0.17 (0.07–0.37) | 0.87 (0.77–0.93) | 0.29 (0.12–0.55) | 0.78 (0.68–0.85) |
| RF = 4 | 0.36 (0.22–0.53) | 0.71 (0.59–0.81) | 0.39 (0.24–0.56) | 0.69 (0.57–0.79) |
| RF = 5 | 0.74 (0.60–0.85) | 0.61 (0.48–0.72) | 0.59 (0.46–0.71) | 0.76 (0.61–0.86) |

KP, predicted keypoints analyzed with a temporal convolutional neural network (TCN); ATT, neural network with attention mechanisms; CF, Commencement of flap & follow through; RF, formation and completion; PPV, positive predictive value; NPV, negative predictive value
entire context in the surgical field to assess skill instead of analyzing instrument motion alone.

Our findings from analysis of predicted instrument tips (KP) in this study reinforce our prior observation that instrument motion in capsulorhexis can be used to discriminate surgical skill [13]. However, performance of algorithms using predicted keypoints was lower than that obtained from using precise manual annotations (AUC 0.79 vs. 0.86). While it is likely that larger datasets can improve accuracy in predicted keypoints, and subsequently in skill assessment, future research should consider methods to encode additional context in the surgical field that is not limited to instrument motion.

Among the feature-based methods, Aug. BoW, a simple method to analyze overall temporal information, appeared to yield a better linear SVM classifier than the other methods. Aug. BoW involves encoding of overall temporal information either at a low- or high-level depending on the type of encoding. On the other hand, DT, SMT, and ApEn involve extracting specific types of temporal features. This specificity limits their utility in a complex real-world environment of the operating room as opposed to a controlled simulated environment in which these methods were developed [32,34]. However, none of the features led to a classifier that was as accurate as the deep learning methods with consistent sensitivity and specificity. Furthermore, large variances in estimates of measures of discrimination observed in our study for the feature-based methods indicate that more data may be necessary to learn a useful classifier with the high-dimensional features. Despite the possibility that larger datasets may result in more accurate feature-based methods, there is little evidence to suggest that they will generalize to new datasets and be useful for surgical skill assessment in the operating room. Moreover, deep learning methods may lead to more potent discriminative classifiers when larger datasets are available.

Performance of algorithms for VBA should be evaluated in the context of the intended application. VBA of surgical skill in the operating room has multiple applications, each with different stakes or consequences. Surgical skill is associated with patient outcomes, therefore, interventions to improve surgeons’ skill can advance quality of care. Skill assessment is key to training surgeons. It supports deliberate practice, and provides summative evaluation at the end of rotations, at the end of each year of training, etc. Some high-stakes applications of surgical skill assessment include certification and re-certification of surgeons for independent practice, determination of operating privileges, and end of career deci-
sions. Not all applications of VBA of surgical skill require the same algorithm performance profile. For example, applications with significant consequences of a false positive, such as certification of surgeons, may require higher specificity (and NPV) than sensitivity (and PPV). In fact, deep learning algorithms in our study, if shown to be externally valid, may not have sufficient specificity for high-stakes assessments, but they may be useful for routine training curricula.

Our study has a few limitations, besides a limited amount of data, that provide directions for future research. Our analyses did not account for multiple videos from the same surgeons, thus, we may have overestimated algorithm performance. Imbalance in distribution of groundtruth labels may have adversely affected algorithm performance. We did not analyze sensitivity of algorithm performance to video resolution. Capsulorhexis is performed through microscopic surgical actions, therefore, videos with a high resolution may enable analysis of granular data patterns. To utilize the learned attention maps for an interpretable assessment of skill and generation of actionable feedback for the surgeon, a qualitative analysis of the usability, feasibility, and correctness of attention maps is necessary. Though our approach uses attention maps to implicitly localize relevant parts of the video without verification, a structured analysis of the acquired attention maps is an interesting direction for future research.

**Conclusion**

Deep learning methods are necessary for VBA of surgical skill in the operating room. While our findings show internal validity of deep learning methods for VBA of surgical skill in capsulorhexis, testing them in additional datasets is necessary to establish their external validity.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11548-022-02681-5.

**Acknowledgements** Dr. Austin Reiter mentored this work in its early stages.

**Funding** Drs. Vedula, Sikder, and Hager are supported by a grant from the National Institutes of Health, USA; NIH 1R01EY033065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**References**

1. Aghdasi N, Bly R, White LW, Hannaford B, Moe K, Lendvay TS (2015) Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. J Surg Res 196(2):302–306
2. Agresti A (2003) Categorical data analysis, vol 482. Wiley, Hoboken
3. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations
4. Bettadapura V, Schindler G, Plötz T, Essa I (2013) Augmenting bag-of-words: data-driven discovery of temporal and structural information for activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2619–2626
5. Birkmeyer JD, Finks JF, O’Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJO (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
6. Carroll N, Richardson I (2016). Software-as-a-medical-device: demystifying connected health regulations. J Syst Inf Technol
7. Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. In: European conference on computer vision, pp 428–441. Springer
8. Deal SB, Stefanidis D, Telem D, Fanelli RD, McDonald M, Michael Ujiki L, Brunt M, Alseidi AA (2017) Evaluation of crowd-sourced assessment of the critical view of safety in laparoscopic cholecystectomy. Surg Endosc 31(12):5094–5100
9. Golnik C, Beaver H, Gauba V, Lee AG, Mayorga E, Palis G, Saleh GM (2013) Development of a new valid, reliable, and internationally applicable assessment tool of residents’ competence in ophthalmic surgery (an American ophthalmological society thesis). Transactions of the American Ophthalmological Society. 111:24
10. Hand DJ, Till RJ (2001) A simple generalisation of the area under the roc curve for multiple classification problems. Mach Learn 45(2):171–186
11. 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
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
13. Kim TS, O’Brien M, Zafar S, Hager GD, Sikder S, Swaroop Vedula S (2019) Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int J Comput Assist Radiol Surg 14(6):1097–1105
14. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. International Conference on Learning Representations, 12
15. Laptev I (2005) On space-time interest points. Int J Comput Vision 64(2–3):107–123
16. Lendvay TS, White L, Kowalewski T (2015) Crowdsourcing to assess surgical skill. JAMA Surg 150(11):1086–1087
17. Lin T, Goyal P, Girshick R, He K, Dollar P (2017) Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 2999–3007
18. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feusner H, Forestier G, Giannarou S, Hashizume M, Katic D, Kenngott H, Kranzfelder M, Malpani A, März K, Neumuth T, Padoy N, Pugh C, Schoch N, Stoyanov D, Taylor R, Wagner M, Hager GD, Jannin P (2017) Surgical data science for next-generation interventions. Nat Biomed Eng 1(9):691–696
19. Malpani SA, Vedula S, Chen CCG, Hager GD (2015) A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. Int J Comput Assist Radiol Surg 10(9):1435–1447
20. Marcano-Cedeno A, Quintanilla-Dominguez J, Cortina-Januchs MG, Andina D (2010) Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In: IECON 2010-36th annual conference on IEEE industrial electronics society, pp 2845–2850. IEEE
21. Pandey VA, Wolfe JHN, Black SA, Cairilos M, Liapis CD, Bergqvist D (2008) Self-assessment of technical skill in surgery: the need for expert feedback. Ann R Coll Surg England 90(4):286–290
22. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf

 Springer
23. Powers MK, Boonjindasup A, Pinsky M, Dorsey P, Maddox M, Su LM, Gettman M, Sundaram CP, Castle EP, Lee JY, Lee BR (2016) Crowdsourcing assessment of surgeon dissection of renal artery and vein during robotic partial nephrectomy: a novel approach for quantitative assessment of surgical performance. J Endourol 30(4):447–452

24. Pugh C, Hashimoto DA, Korndorffer Jr JR (2020) The what? how? and who? of video based assessment. Am J Surg

25. Robertson S (2004) Understanding inverse document frequency: on theoretical arguments for idf. J Doc

26. Sharma Y (2014) Surgical skill assessment using motion texture analysis. PhD thesis, Georgia Institute of Technology

27. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5693–5703

28. Valanci-Aroesty S, Alhassan N, Feldman LS, Landry T, Mas-tropietro V, Fiore Jr J, Lee L, Fried GM, Mueller CL (2020) Implementation and effectiveness of coaching for surgeons in practice—a mixed studies systematic review. J Surg Educ

29. 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

30. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057

31. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp 207–212

32. Zia A, Essa I (2018) Automated surgical skill assessment in rmis training. Int J Comput Assist Radiol Surg 13(5):731–739

33. Zia A, Sharma Y, Bettadapura V, Sarin EL, Clements MA, Essa I (2015) Automated assessment of surgical skills using frequency analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp 430–438. Springer

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

35. Zia A, Sharma Y, Bettadapura V, Sarin EL, Ploetz T, Clements MA, Essa I (2016) Automated video-based assessment of surgical skills for training and evaluation in medical schools. Int J Comput Assist Radiol Surg 11(9):1623–1636

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.