Noname manuscript No.
(will be inserted by the editor)

Objective Assessment of Intraoperative Technical Skill in Capsulorhexis using Videos of Cataract Surgery

Tae Soo Kim · Molly O'Brien · Sidra Zafar · Gregory D. Hager · Shameema Sikder* · S. Swaroop Vedula*.

Received: date / Accepted: date

Abstract Purpose Objective assessment of intraoperative technical skill is necessary for technology to improve patient care through surgical training. Our objective in this study was to develop and validate deep learning techniques for technical skill assessment using videos of the surgical field.

Methods We used a data set of 99 videos of capsulorhexis, a critical step in cataract surgery. One expert surgeon annotated each video for technical skill using a standard structured rating scale, the International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric:phacoemulsification (ICO:OSCAR-phaco). Using two capsulorhexis indices in this scale (commencement of flap & follow through; formation and completion), we specified an expert performance when at least one of the indices was 5 and the other index was at least 4, and novice, otherwise. In addition, we used scores for capsulorhexis commencement and capsulorhexis formation as separate ground truths (Likert scale of 2 to 5; analyzed as 2/3, 4, and 5). We crowdsourced annotations of instrument tips. We separately modeled instrument trajectories and optical flow using temporal convolutional neural networks to predict a skill class (expert/novice) and score on each item for capsulorhexis in ICO:OSCAR-phaco. We evaluated the algorithms in a 5-fold cross-validation and computed accuracy and area under the receiver operating characteristics curve (AUC).

Results The accuracy and AUC were 0.848 and 0.863 for instrument tip velocities, and 0.634 and 0.803 for optical flow fields, respectively.

Conclusions Deep neural networks effectively model surgical technical skill in capsulorhexis given structured representation of intraoperative data such

* authors equally contributed to this work

S. Swaroop Vedula
3400 N. Charles Street, Malone Hall 340, Baltimore, MD 21218
E-mail: vedula@jhu.edu

as optical flow fields extracted from video or crowdsourced tool localization information.

Keywords Surgical skill assessment · Neural networks · Deep learning · Capsulorhexis · Cataract surgery · Tool trajectories · Crowdsourcing

1 Introduction

Cataract surgery is a definitive intervention to improve visual impairment caused by an aging lens. Although cataract surgery is one of the most commonly performed procedures across the world, acquiring skill and competency with the procedure is not trivial. In fact, the learning curve for cataract surgery is quite steep; surgical skill continues to improve significantly well beyond the first 80 resident cases [14]. For instance, data from the United States show that about 14% of all graduating ophthalmology residents reported insufficient skill in cataract surgery, and favored additional training [11]. Despite being a critical skill that ophthalmologists across the globe should efficiently acquire, training in cataract surgery is supported only by manual assessments using heterogeneous structured rating scales [13]. Data science and machine learning techniques applied to readily accessible videos of the surgical field hold substantial potential for automated objective assessment of technical skill in cataract surgery. However, previous research in data science for cataract surgery emphasized automated detection of surgical phase [18,19]. Our objective in this study is to develop and validate deep learning algorithms for objective assessment of technical skill in capsulorhexis. Together with phacoemulsification, capsulorhexis is one of the critical steps in cataract surgery, difficult to learn and teach [11], and may affect surgical outcomes.

Data-driven methods for objective assessment of surgical technical skill have been explored in multiple surgical disciplines, with some success, and mostly in the context of simulation [17]. Recent technical advances, particularly in deep learning, are transforming algorithms to automate understanding human activities in surgical videos [1,3,4]. Despite these advances, the pace of progress is limited by the scale of data that may be captured in the operating room (OR). In their current form, data sets for assessment of technical skill in the OR are insufficient to train modern deep networks in a purely data-driven fashion. To properly harness the potential of such powerful yet data-dependent learning models for fine-grained tasks under data scarce settings, we believe it is essential to identify appropriate structure in the problem for the model to exploit. For example, in cataract surgery, instrument usage information is a strong indicator of surgical context [1,2]. Applying the structured prior knowledge to identifying surgical phase by recognizing tool usage, deep neural network based approach shows excellent phase recognition performance [18]. For the task of identifying technical surgical skill from video, we hypothesize that learning a model on motion based data representation is more effective over raw RGB inputs. Working with motion based representation such as optical flow or trajectories provides robustness to discrepencies that may exist

across institutes and to intra sample visual variances. In this work, we study the effects of varying amounts of structure in data on neural network's ability to recognize surgical technical skill in capsulorhexis.

In summary, major contributions in this work are *a*) techniques to automate objective assessment of technical skill assessment in capsulorhexis at scale, and *b*) comparative evaluation of algorithm performance with different amounts of structure in time series intraoperative video images.

2 Methods

We describe our technical approach using temporal convolutional neural networks (TCNs) [9] to objectively assess intraoperative technical skill using videos of capsulorhexis. Videos of the surgical site serve as rich sources of information containing surgical context, motions and intraoperative actions. We model surgical videos as a time-series of local spatio-temporal representations. By end-to-end learning of technical skill with TCNs, we wish to jointly learn discriminative local patterns in video as well as temporal dependencies of such patterns. We describe our TCN approach in Section 2.1. Instead of directly observing RGB pixel values of videos, we hypothesize that motion captures more scene structure relevant to surgical skill. In Section 2.2, we explore representing local video-snippets as optical flow fields computed from state-of-the-art optical flow estimation methods such as [15]. In capsulorhexis and surgery in general, surgeons interact with the surgical site through instruments. Then for the task of assessing the surgeon's technical skill, we believe that surgical tool movement during surgery encodes the most relevant information for identifying technical skill. Therefore, we explore learning skill models with tool trajectories. We describe our skill assessment approach with tool trajectories in Sections 2.3 and 2.4.

2.1 TCNs for Skill Assessment

We model a video X as a temporal concatenation of N local representations such that

$$X = \{\overrightarrow{x_1}, \cdots, \overrightarrow{x_N}\}, \quad \overrightarrow{x_n} \in \mathbb{R}^m, X \in \mathbb{R}^{N \times m} \tag{1}$$

where $\overrightarrow{x_n}$ is the n-th local spatio-temporal encoding of m dimensions. For a TCN, which maps X, to a skill label, y, with L layers, each layer l has F_l convolutional filters of size $d \times F_{l-1}$ where d is the temporal filter length. During a forward pass, given a signal from the previous layer, $h^{(l-1)}$, the l-th layer of TCN computes the activation $h^{(l)}$ as

$$h^{(l)} = \sigma(W_l * h^{(l-1)}) \tag{2}$$

where W_l is the set of filters at the l-th layer, $\sigma(\cdot)$ is a non-linearity (Rectified Linear Unit) and $*$ is a convolution operator. A global average pooling layer in the temporal domain as in [7] ensures that regardless of the length of the

trajectory, the TCN always outputs a fixed size representation, $h^{(L)} \in \mathbb{R}^{F_L}$. Finally, TCN includes a linear layer, $W_C \in \mathbb{R}^{C \times F_L}$ and the corresponding bias term b_C, followed by a softmax activation to compute the skill class probabilities, where C is the number of skill labels.

$$\hat{y} = softmax(W_C h^{(L)} + b_C) \tag{3}$$

The presented TCN approach shares many similarities with the approach proposed in [7,9].

2.2 Local Representation using Optical Flow Fields

To model local spatial-temporal movements from RGB videos, we utilize optical flow fields. For the TCN model that learns skill from optical flow fields, a set of frames in the n-th local window, $\{X_t \cdots X_{t+\delta}\}_n$, is encoded into a feature representation as:

$$\begin{aligned} \Delta I_n &= G(X_t, X_{t+\delta}), \;\; \Delta I_n \in \mathbb{R}^{W \times H \times 2} \\ \overrightarrow{x_n} &= f(\Delta I_n), \;\; \overrightarrow{x_n} \in \mathbb{R}^{2WH} \end{aligned} \tag{4}$$

where ΔI_n is the estimated optical flow fields of size W by H by 2, G is an optical flow computation model, f is a flattening operator that rasterizes the flow fields in to a 1-dimensional vector and $(X_t, X_{t+\delta})$ is the input image pair. In our experiments, we apply the model described in [15] to compute optical flow fields; however, the choice of G remains flexible.

2.3 Local Representation with Structured Data: Tool Trajectories

Under the hypothesis that more structured data such as intraoperative tool movements provide the right structured information for learning skill, we collect tool trajectories from the crowd using Amazon Mechanical Turk framework [10]. The data collection procedure follows the recent work of [6] which validates the effectiveness of collecting tool tip trajectories in cataract surgery at scale. Briefly, crowd workers are asked to annotate the tips of the tools. If the tool tips are not in the field of view but the instrument is in the eye, the crowd workers are asked to annotate the points on the tools that are closest to the tips.

Similar to the local optical flow based representation discussed in 2.2, a set of frames in the n-th local window, $\{X_t \cdots X_{t+\delta}\}_n$, is encoded using crowd annotations:

$$\overrightarrow{x_n} =< p_t^1(x), p_t^1(y), p_t^2(x), p_t^2(y), \cdots, p_t^d(x), p_t^d(x) >, \;\; \overrightarrow{x_n} \in \mathbb{R}^{2d} \tag{5}$$

where d is the number of defined tool tip locations and $(p_t^d(x), p_t^d(y))$ is the pixel location of the d-th tool tip at time t. In this work, tool trajectory includes three tool keypoints: the tip of the cystotome and the two tips of the Utrata forceps resulting in a 6 dimensional representation per timestep.

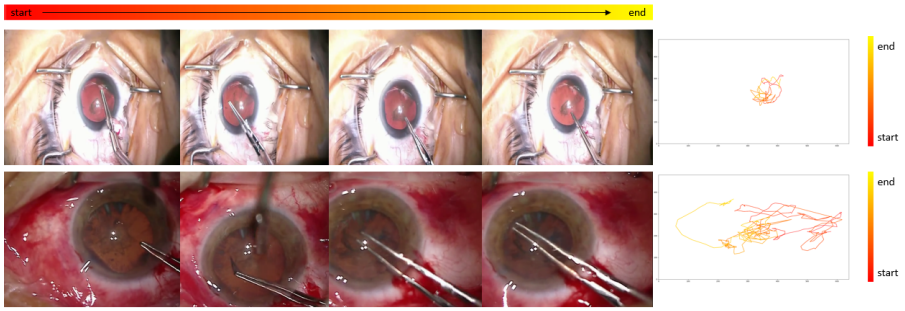

Fig. 1: Trajectory of tools in capsulorhexis. The distortion in forcep trajectory shown in the bottom row illustrates that there exists significant trajectory appearance variance among trials arising from characteristics of intraoperative surgical data such as surgical site movement. The top row shows a sample where the surgical site is relatively stationary, leading to a clean circular trajectory. Additionally, the two samples vary in scale due to different data capture settings. Red encodes the beginning portion of the procedure; yellow corresponds to the ending stages of capsulorhexis. Best viewed in color.

2.4 Tool Trajectory Representation for Effective Skill Assessment

With intraoperative surgical data such as tool trajectories, it is often the case that there exists significant variance among trials. For example in capsulorhexis, the circular tearing motion using the forceps may not look circular at all when the surgical site itself translates. Figure 1 illustrates such variability common in intraoperative surgical data. In both samples, the forceps are creating a circular tear but the resulting trajectory shape is dramatically different between the cases. Moreover, in Figure 1, the scale difference among samples is also evident. When the goal of the model is to identify tool motions that indicate technical surgical skill, such characteristics of the data pose a significant hurdle for the learning task. Compounded by the additional difficulty that intraoperative surgical data with skill information is often limited in scale, finding an effective representation of the data for surgical skill assessment model learning is crucial.

We observe that tool trajectory representation with pixel locations is vulnerable to many compounding factors commonly found in surgical data such as intraoperative surgical site movement, variations in data collection settings and inter-site tool differences. Hence, instead of using tool tip positions, we experiment with representing trajectories as a time-series of tool tip velocities. More specifically, compared to the representation presented in Section 2.3, the local n-th window, $\overrightarrow{x_n}$, is now defined as:

$$\begin{aligned} \overrightarrow{p_n} &= < p_t^1(x), p_t^1(y), \cdots, p_t^d(x), p_t^d(x) > \quad , \quad \overrightarrow{p_n} \in \mathbb{R}^{2d} \\ \overrightarrow{p_{n+1}} &= < p_{t+\delta}^1(x), p_{t+\delta}^1(y), \cdots, p_{t+\delta}^d(x), p_{t+\delta}^d(x) > \\ \overrightarrow{x_n} &= \overrightarrow{p_{n+1}} - \overrightarrow{p_n} \quad , \quad \overrightarrow{x_n} \in \mathbb{R}^{2d} \end{aligned} \tag{6}$$

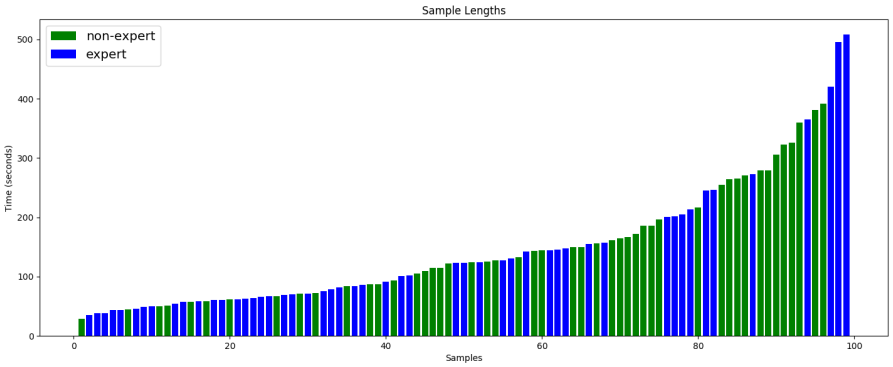

Fig. 2: Duration of capsulorhexis in our sample. Mean (SD) is 147.97 (104.81) seconds. The green bar represents a sample from a novice class and blue otherwise. Best viewed in color.

Compared to raw tool trajectories presented in Section 2.3, $\overrightarrow{x_n}$ now encodes the first derivative information which is more robust to common data complication scenarios mentioned above. We argue that such simple encoding of trajectories is more effective for the task of surgical skill assessment and can improve results.

3 Experiments

3.1 Data representations

We evaluated several data representations including tool tip positions (TP), tool tip velocities (TV), optical flow fields (FF), as well as augmented representations including FF + TP, FF + TV, and FF + TP + TV. We hypothesized that concatenating different data representations may have complementary information and thereby, improve algorithm performance.

3.2 Data set

Our data set comprised of 99 videos of capsulorhexis captured from the operating microscope. We processed the videos to a 640 by 480 resolution at 59 frames per second. One expert surgeon watched each video, with no information on identity of the operating surgeon, and evaluated for technical skill using a previously validated structured rating scale - the International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric:Phacoemulsification (ICO-OSCAR:phaco) [13]. This rubric includes two items, one for each major activity in capsulorhexis —*commencement of flap & follow-through* (CF) score and *formation and completion* (RF). Both

Table 1: Data set statistics for cross validation folds. We chose a partition of the data set that minimized differences in within-fold sum of durations across folds with balanced distribution of expert and non-expert instances across folds. CF: commencement of flap & follow-through. RF: rhexis formation

	Fold 1	Fold 2	Fold3	Fold 4	Fold 5
Expert	10	10	10	10	11
Non-Expert	9	9	9	9	12
CF=2 or 3	4	1	3	4	2
CF=4	8	9	6	6	10
CF=5	7	9	10	9	11
RF=2 or 3	6	4	5	5	3
RF=4	3	6	6	6	12
RF=5	10	9	8	8	8
Total Duration (s)	2950.9	2909.39	2934.17	2926.43	2927.74

items are assessed on a Likert scale of 2 to 5, with a larger score indicating better skill.

For our analyses, we specified the ground truth as a binary (expert/non-expert) skill class label, in addition to scores on each of the two items in ICO-OSCAR:phaco for capsulorhexis. We specified a given video as an expert instance when it received a score of 5 on at least one item for capsulorhexis and at least 4 on the other item. We did not use faculty/trainee appointment status as a surrogate for expert/non-expert skill class labels because it is helpful neither for the educators (to give effective feedback) nor for the trainees (for deliberate practice). Table 1 illustrates number of instances of capsulorhexis by the ground truth binary skill class label. Figure 2 illustrates the distribution of duration of capsulorhexis in our sample.

3.3 Experimental Setup

We conducted two sets of experiments. In the first, we predicted a binary technical skill class label (expert/novice), and in the second, we jointly predicted the binary skill class and scores on individual items for capsulorhexis in ICO-OSCAR:phaco (CF and RF). That is, given a sample, each prediction incurred a loss and we jointly optimized the model to maximize accuracy for all three predictions. In addition, we treated the second experiment as a classification task with three classes —a score of 2 or 3, a score of 4, and a score of 5. We chose this approach owing to very few samples in our data set with a ground truth score of 2 for CF and RF.

We performed 5-fold cross validation using the same data partitions for all experiments in this study. Retaining one fold as the test, we iteratively

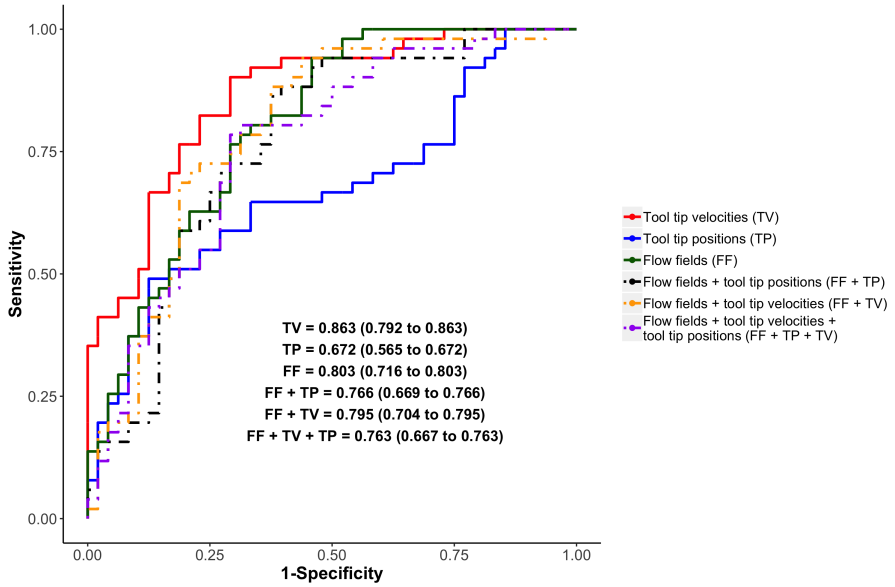

Fig. 3: Receiver operating characteristic curves for algorithms for technical skill assessment in capsulorhexis.

used three of the remaining four folds for training the TCN and one fold for validation.

During training, after every training epoch, we computed accuracy on the validation set and stored the state of the model if it improved. We did not use data from the test set for model development. Within each cross validation partition, we used the average of the class probabilities across the four validated models. In addition, our experimental set up heavily penalizes overfitted models that fail to generalize to unseen samples.

Model Settings: We used identical learning parameters and TCN settings for all reported experiments. Figure 4 depicts our TCN architecture. All convolution layers used 1D temporal convolution filters of length 8 and max pooling layers downsample the input by a factor of 2 in the temporal dimension. We optimized the model for 50 epochs using ADAM [8], with a learning rate of 0.001 throughout and with a L2 weight decay term of magnitude 1e-5. Standard neural network techniques often found effective in the computer vision literature such as dropout [16] and batch normalization [5] did not exhibit meaningful differences empirically. We implemented our model using PyTorch (0.4.1) [12] with Python 3.6 (implementation code may be publicly available in the future). For experiments using optical flow, the extracted flow fields are of dimensions $W = 10$ by $H = 7$ by 2 resulting in a flattened representation

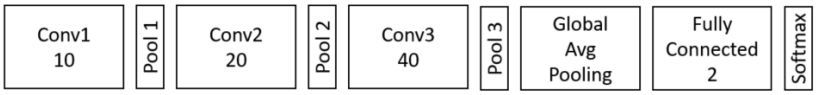

Fig. 4: Skill Assessment TCN architecture. TCN architecture has 3 temporal convolution layers, 3 max pool layers, global average pooling layer followed by a fully conntected linear layer and a softmax activation. Numbers of temporal filters are denoted in the box. All convolutional filters are of length 8 and max pool reduces the temporal dimension by a factor of 2.

Table 2: Skill classification accuracy per fold and across folds (numbers in parentheses are 95% confidence intervals). FF = Flow fields; TP = Tool tip positions; TV = Tool tip velocities.

Representation	Fold 1	Fold 2	Fold3	Fold 4	Fold 5	Accuracy Across Folds
TV	0.789	0.737	**0.842**	**1.000**	**0.870**	**0.848** (0.770 to 0.926)
TP	0.526	0.632	0.632	0.789	0.565	0.629 (0.579 to 0.679)
FF	0.684	0.684	0.474	0.632	0.696	0.634 (0.561 to 0.707)
FF + TP	0.789	0.684	0.632	0.632	0.696	0.686 (0.636 to 0.736)
FF + TV	0.737	**0.789**	0.632	0.632	0.696	0.696 (0.644 to 0.751)
FF + TP + TV	**0.842**	0.632	0.684	0.684	0.696	0.708 (0.646 to 0.770)

of 140 dimensions. Further details on collection of tool tip data are described in [6]. We computed estimates and 95% confidence intervals (CIs) for global metrics including accuracy and area under the receiver operating characteristic curve (AUC), in addition to class-specific sensitivity and specificity.

3.4 Results

Table 2 illustrates skill classification (expert/novice) accuracy of the algorithm for data representations described in Section 2, within and across cross validation folds. The overall accuracy was highest with TV. While combining data representations improved accuracy of the algorithm, their estimates were lower than that with TV. This difference is likely because representations that are sensitive to spatial position such as TP and FF poorly generalize, especially with limited training data. On the other hand, TV is robust to variation in spatial position of the tool tip across videos, and thereby, allowing the model to effectively generalize.

Our estimates in Tables 2 and 3 show that algorithms trained with different data representations vary in their performance characteristics. TV yielded an algorithm that had the best sensitivity but lower, albeit still one of the highest, specificity, along with high AUC. The TCN algorithm with TP alone had the lowest performance. Combining FF with TP adversely affected sensitivity

Table 3: Estimates of performance for assessing a global technical skill class (expert/novice) in capsulorhexis (95% confidence intervals in parentheses). FF = Flow fields; TP = Tool tip positions; TV = Tool tip velocities; AUC = Area under the receiver operating characteristic curve.

Encoding	Sensitivity	Specificity	AUC
TV	0.824 (0.697 to 0.904)	0.708 (0.568 to 0.818)	0.863 (0.792 to 0.863)
TP	0.647 (0.510 to 0.764)	0.521 (0.383 to 0.655)	0.672 (0.565 to 0.672)
FF	0.725 (0.591 to 0.829)	0.708 (0.568 to 0.818)	0.803 (0.716 to 0.803)
FF + TP	0.667 (0.530 to 0.780)	0.750 (0.612 to 0.851)	0.766 (0.669 to 0.766)
FF + TV	0.745 (0.611 to 0.845)	0.708 (0.568 to 0.818)	0.795 (0.704 to 0.795)
FF + TP + TV	0.725 (0.591 to 0.829)	0.708 (0.568 to 0.818)	0.763 (0.667 to 0.763)

obtained with FF alone, but improved estimated specificity. Combined TV with FF yielded comparable performance with FF alone.

Tables 4 and 5 show estimates of performance to predict scores on individual items in capsulorhexis (CF and RF). Accuracy to classify expert/novice with the jointly trained algorithm was comparable to that described in Table 2. However, performance of the algorithm was heterogeneous across classes representing scores on CF and RF. We observed high specificity for score 2/3 and high sensitivity for score 5, suggesting that our findings are strongly influenced by the score imbalance for CF and RF in our data set. It is unlikely that the model has successfully learned an adequate representation for the severely under-represented classes. However, confusion matrix analysis presented in Table 6 illustrates that the model is less likely to make errors between the extreme skill ratings. Presumably, misclassifying a CF-5 sample as a CF-4 is a less serious error than as a CF-2/3. We show that such confusion never happens in the converged models. Moreover, for both CF and RF evaluations, model misclassifies 2/3 samples as skill rating 4 more often than as a 5. Given such model behavior, we conjecture that the presented model has learned meaningful representation for this task but it is still not adequate enough for operational use.

4 Discussion

In this study evaluating deep neural networks for objective assessment of technical skill in capsulorhexis, our experiments using videos of the surgical field yielded algorithms with a high AUC to predict a binary class label (expert/novice). Of all the data representations we evaluated, algorithms using TV, a more structured representation than TP or FF, had the best performance. Finally, a model trained to jointly predict a binary skill class and scores on individual items, CF and RF, performed poorly, likely owing to an imbalanced data set.

Table 4: Estimated accuracy of the TCN using tool tip velocities (TV) to jointly predict a binary skill class label and scores on individual items for capsulorhexis in ICO-OSCAR:phaco. Numbers in parentheses are 95% confidence intervals; standard deviation reported for micro averaged accuracy. CF: commencement of flap & follow-through. RF: formation and completion.

Score Item	Accuracy Across Folds
Expert / Non-Expert	0.848 (0.796 to 0.900)
CF all - micro	0.657 (0.559 to 0.743)
CF all - macro	0.771 ± 0.091
CF 2/3	0.859 (0.777 to 0.914)
CF 4	0.677 (0.580 to 0.761)
CF 5	0.778 (0.686 to 0.848)
RF all - micro	0.525 (0.428 to 0.621)
RF all - macro	0.684 ± 0.076
RF 2/3	0.717 (0.622 to 0.796)
RF 4	0.596 (0.497 to 0.687)
RF 5	0.737 (0.643 to 0.814)

Table 5: Estimates of TCN performance for score on individual items to assess capsulorhexis in ICO-OSCAR:phaco (95% confidence intervals in parentheses). FF = Flow fields; TP = Tool tip positions; TV = Tool tip velocities; AUC = Area under the receiver operating characteristic curve. CF: Commencement of flap & follow through; RF: formation and completion

Encoding	Sensitivity	Specificity	AUC
CF = 2/3	0 (0 to 0.215)	1 (0.957 to 1)	0.761 (0.689 to 0.801)
CF = 4	0.641 (0.484 to 0.773)	0.700 (0.575 to 0.801)	0.815 (0.746 to 0.870)
CF = 5	0.870 (0.743 to 0.939)	0.698 (0.565 to 0.805)	0.805 (0.744 to 0.860)
RF = 2/3	0.087 (0.024 to 0.268)	0.908 (0.822 to 0.955)	0.746 (0.662 to 0.791)
RF = 4	0.455 (0.298 to 0.620)	0.667 (0.547 to 0.768)	0.747 (0.665 to 0.814)
RF = 5	0.814 (0.674 to 0.903)	0.679 (0.548 to 0.786)	0.768 (0.694 to 0.831)

Table 6: Confusion matrices for individual item score prediction for both CF and RF. CF: Commencement of flap & follow through; RF: formation and completion; GT: Ground truth label; Pred.: Model prediction.

CF	Pred. 2/3	Pred. 4	Pred. 5
GT 2/3	0	12	2
GT 4	0	25	14
GT 5	0	6	40

RF	Pred. 2/3	Pred. 4	Pred. 5
GT 2/3	2	14	7
GT 4	7	15	11
GT 5	0	8	35

Our findings suggest that structured data representations such as tool tip trajectories yield better performing models. We relied on crowdsourced annotations of the surgical tool tips to compute velocities. Data may also be structured in other ways to register or normalize tool trajectories. While videos of the surgical field are readily available in minimally invasive and microsurgery,

automated methods are necessary to structure these data at scale. Recent research to identify instruments in cataract surgery suggests that it is feasible to segment the instrument and extract structured data for skill assessment [18].

Our work extends the automated skill assessment paradigm beyond predicting a binary class (expert/novice) to a granular evaluation of individual items on a standardized rating scale. Item-specific assessments can make these automated methods more relevant to surgical educators and training curricula. For data-driven interventional healthcare to truly make an impact on improving the quality and efficiency of care, we believe that surgical technical skill assessment models should be human-interpretable, i.e., explain predictions in terms of events or measures that educators and trainees can comprehend and use to understand their performance,and generate targeted feedback for users to improve their performance. However, item-specific assessments are interpretable only to the extent of rigor with which the ground truth rating scales are developed.

Limitations in this study help identify areas for further research. Our data set had few instances with scores of 2/3 for CF and RF. Reference data sets that are balanced for skill measures and representative of surgeons across the skill spectrum are necessary to advance automation of surgical technical skill assessments. While our data set include ratings from one surgeon, reference data sets should include ground truth specified by multiple raters to establish inter-rater reliability.

In conclusion, our study led to validation of a deep neural network to assess technical skill in capsulorhexis using crowdsourced annotations of instrument tips in videos of the surgical field.

Acknowledgements Dr. Anand Malpani advised on crowdsourcing for annotation of instruments and Adit Murali supported cleaning the data.

5 Compliance with Ethical Standards

Funding: This study was supported by funds from the Wilmer Eye Institute Pooled Professor's Fund (PI: Dr. Sikder), an unrestricted research grant to the Wilmer Eye Institute from Research to Prevent Blindness, and a research grant from The Mitchell Jr. Trust (PI: Dr. Sikder).
Conflict of Interest: The authors declare that they have no conflict of interest.
Ethical Approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional review board and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.
Informed Consent: Informed consent was obtained from all individual participants included in the study.

References

1. Bouget, D., Allan, M., Stoyanov, D., Jannin, P.: Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Medical Image Analysis **35**,

633–654 (2017)
2. Bouget, D., Lalys, F., Jannin, P.: Surgical Tools Recognition and Pupil Segmentation for Cataract Surgical Process Modeling. In: Medicine Meets Virtual Reality - NextMed, vol. 173, pp. 78–84. IOS press books, Newport beach, CA, United States (2012). URL http://www.hal.inserm.fr/inserm-00669660
3. Du, X., Kurmann, T., Chang, P.L., Allan, M., Ourselin, S., Sznitman, R., Kelly, J.D., Stoyanov, D.: Articulated multi-instrument 2d pose estimation using fully convolutional networks. IEEE Transactions on Medical Imaging **1**(1), 99 (2018)
4. Gao, Y., Vedula, S.S., Reiley, C., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Bejar, B., Yuh, D.D., Chen, C., Vidal, R., Khudanpur, S., Hager, G.D.: The jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: In Modeling and Monitoring of Computer Assisted Interventions (M2CAI), MICCAI (2014)
5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: F.R. Bach, D.M. Blei (eds.) ICML, *JMLR Workshop and Conference Proceedings*, vol. 37, pp. 448–456. JMLR.org (2015)
6. Kim, T.S., Malpani, A., Reiter, A., Hager, G.D., Sikder, S., Vedula, S.S.: Crowdsourcing annotation of surgical instruments in videos of cataract surgery. In: Medical Image Computing and Computer Assisted Intervention LABELS Workshop (MICCAI-LABELS), pp. 121–130 (2018)
7. Kim, T.S., Reiter, A.: Interpretable 3d human action analysis with temporal convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 1623–1631 (2017)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. 3rd International Conference for Learning Representations, San Diego (2015). URL http://arxiv.org/abs/1412.6980; accessed on January 26, 2019
9. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: CVPR (2017)
10. Little, G., Chilton, L.B., Goldman, M., Miller, R.C.: Turkit: Tools for iterative tasks on mechanical turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, HCOMP '09, pp. 29–30. ACM, New York, NY, USA (2009). DOI 10.1145/1600150.1600159. URL http://doi.acm.org/10.1145/1600150.1600159
11. McDonnell, P.J., Kirwan, T.J., Brinton, G.S.: Perceptions of recent ophthalmology residency graduates regarding preparation for practice. Ophthalmology **114**(2), 387–391 (2007)
12. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
13. Puri, S., Srikumaran, D., Prescott, C., Tian, J., Sikder, S.: Assessment of resident training and preparedness for cataract surgery. Journal of Cataract and Refractive Surgery **43**(3), 364–368 (2017)
14. Randleman, J., Wolfe, J.D., Woodward, M., Lynn, M.J., Cherwek, D., Srivastava, S.K.: The resident surgeon phacoemulsification learning curve. Archives of Ophthalmology **125**(9), 1215–1219 (2007)
15. Ranjan, A., Black, M.J.: Optical flow estimation using a spatial pyramid network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
16. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014). URL http://jmlr.org/papers/v15/srivastava14a.html
17. Vedula, S.S., Ishii, M., Hager, G.D.: Objective assessment of surgical technical skill and competency in the operating room. Annual Review of Biomedical Engineering **19**(1), 301–325 (2017). DOI 10.1146/annurev-bioeng-071516-044435. URL https://doi.org/10.1146/annurev-bioeng-071516-044435. PMID: 28375649
18. Zisimopoulos, O., Flouty, E., Luengo, I., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D.: Deepphase: Surgical phase recognition in cataracts videos. In: MICCAI (2018)
19. Zisimopoulos, O., Flouty, E., Stacey, M., Muscroft, S., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D.: Can surgical simulation be used to train detection and classification neural networks? Healthcare Technology Letters **4**(5), 216–222 (2017)