tvst

Article

PhacoTrainer: A Multicenter Study of Deep Learning for Activity Recognition in Cataract Surgical Videos

Hsu-Hang Yeh[1], Anjal M. Jain[2], Olivia Fox[3], and Sophia Y. Wang[1,2]

[1] Department of Biomedical Data Science, Stanford University, Palo Alto, CA, USA
[2] Department of Ophthalmology, Byers Eye Institute, Stanford University, Palo Alto, CA, USA
[3] Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, MD, USA

Correspondence: Sophia Y. Wang, Department of Biomedical Data Science, Stanford University, 2370 Watson Court, Palo Alto, CA 94303, USA.
e-mail: sywang@stanford.edu

Received: August 5, 2021
Accepted: October 5, 2021
Published: November 16, 2021

Keywords: artificial intelligence; deep learning; cataract surgery; surgical training

Citation: Yeh HH, Jain AM, Fox O, Wang SY. PhacoTrainer: A multicenter study of deep learning for activity recognition in cataract surgical videos. Transl Vis Sci Technol. 2021;10(13):23, https://doi.org/10.1167/tvst.10.13.23

Purpose: To build and evaluate deep learning models for recognizing cataract surgical steps from whole-length surgical videos with minimal preprocessing, including identification of routine and complex steps.

Methods: We collected 298 cataract surgical videos from 12 resident surgeons across 6 sites and excluded 30 incomplete, duplicated, and combination surgery videos. Videos were downsampled at 1 frame/second. Trained annotators labeled 13 steps of surgery: create wound, injection into the eye, capsulorrhexis, hydrodissection, phacoemulsification, irrigation/aspiration, place lens, remove viscoelastic, close wound, advanced technique/other, stain with trypan blue, manipulating iris, and subconjunctival injection. We trained two deep learning models, one based on the VGG16 architecture (VGG model) and the second using VGG16 followed by a long short-term memory network (convolutional neural network [CNN]– recurrent neural network [RNN] model). Class activation maps were visualized using Grad-CAM.

Results: Overall top 1 prediction accuracy was 76% for VGG model (93% for top 3 accuracy) and 84% for the CNN–RNN model (97% for top 3 accuracy). The microaveraged area under receiver-operating characteristic curves was 0.97 for the VGG model and 0.99 for the CNN–RNN model. The microaveraged average precision score was 0.83 for the VGG model and 0.92 for the CNN–RNN model. Class activation maps revealed the model was appropriately focused on the instrumentation used in each step to identify which step was being performed.

Conclusions: Deep learning models can classify cataract surgical activities on a frame-by-frame basis with remarkably high accuracy, especially routine surgical steps.

Translational Relevance: An automated system for recognition of cataract surgical steps could provide to residents automated feedback metrics, such as the length of time spent on each step.

Introduction

Cataract is a clouding of the lens which causes poor vision and is the leading cause of blindness worldwide.[1] Cataract surgery restores vision by replacing the cloudy cataractous lens with a clear lens implant and is the most commonly performed surgery in the United States.[2] Cataract surgery is a highly delicate and challenging operation involving manipulations of micrometer-thick tissues under microscopic magnification.[3] Although some cases may be relatively straightforward, complex cases may include the management of poor dilation, floppy iris syndrome, zonular weakness, or advanced cataracts, the difficulty of which may lead to complications such as vitreous loss. The safe and skilled performance of cataract surgery is a primary goal of residency training in ophthalmology.

The mainstay of surgical training is case-by-case, one-on-one, real-time feedback from attending preceptors to trainees. Videos of cataract surgery are also captured routinely by trainees to review subsequently for complications or learning opportunities. However, human rating systems[4] for video performance are manual and time consuming, and infeasible for use on a large scale (e.g., for every recorded surgery). Thus, aside from the time in the operating room itself, there are limited significant opportunities for external feedback for one's personal surgical performance. Furthermore, it is difficult to tabulate efficiently objective metrics for cataract surgical performance, such as the length of time spent on certain important steps of cataract surgery such as capsulorrhexis or phacoemulsification. Other metrics of surgical training such as the use of advanced techniques for complex cases and rates of complications requiring anterior vitrectomy can only be tabulated manually by the surgeon, which can be onerous.

Motion analysis using traditional computer vision techniques has been used to provide objective measures of surgical performance in cataract surgeries.[5] These analyses could distinguish between novice and expert surgeons in instrument movement and path. However, they do not distinguish between different steps of cataract surgeries.

The development of algorithms that can recognize both basic and complex activities in cataract surgery would enable detailed yet automated analyses of cataract surgical videos, including the time spent on individual steps and the presence of complications and nonroutine surgical activities. This information could be used to summarize cataract surgery experience and performance, which would be particularly useful to trainees seeking to further improve their surgical performance.

Advances in artificial intelligence and computer vision in medicine now provide an opportunity to analyze video data on a large scale, for example, with recently developed algorithms for detecting patient mobilization in intensive care units[6] and for detecting activities and instruments in general surgery, enabling automated surgical skills assessment in that domain.[7] Deep learning has also been successfully applied to cataract surgery analysis.[8–10] However, most studies used videos from a single institute and did not incorporate advanced surgical steps, thus limiting their generalizability.

Our goal was to train a deep neural network to automatically recognize the various steps of cataract surgery, including both routine steps as well as the use of advanced techniques for complex surgeries such as use of the malyugin ring, staining with trypan blue, and the management of complications, including anterior vitrectomy.

Methods

Data Source and Preprocessing

We collected a sample of 298 resident cataract surgical videos that had been routinely recorded during the residency training of 12 surgeons across 6 different sites. All data were deidentified and this research was deemed to be exempt by the institutional review board. The original footage was in 1080 × 1920 resolution, which was downsampled to a resolution of 256 × 456. Videos that were significantly incomplete (e.g., recording began after main wound creation), duplicated, or contained steps of another major surgery (e.g., phacotrabeculectomy) were excluded ($n = 30$), leaving 268 videos in the final dataset, which amounted to approximately 131 hours of included footage. Videos of poor quality caused by surgeons' inexperience, such as videos with decentered or out of focus views, were kept in the dataset.

A team of four trained annotators used VIA software[11] to manually label the start and end times for 13 specific steps of cataract surgery: create wound, injecting substance into the eye, capsulorrhexis, hydrodissection, phacoemulsification, irrigation/aspiration, place lens (including axis marking for toric lenses), remove viscoelastic, close wound, advanced technique/other (including anterior vitrectomy, placement of capsular support devices, limbal relaxing incisions, superior rectus traction suture, conjunctival vessel cautery, and vitreous trimming at wound), staining with trypan blue, manipulation of iris (including placement and removal of malyugin rings/iris hooks, and repositioning prolapsed iris), injection into the subconjunctival space (including sub-Tenon's blocks). The frames containing none of these activities are labeled as "no label," resulting in a total of 14 classes of labels. The start time was marked when the relevant instruments are seen on the screen, and the end time when the instruments leave the screen. By this definition, failed attempts are also labeled as within the step. All annotators received a presentation on the steps of cataract surgery and observed several examples of cataract surgical videos being annotated in real-time by the ophthalmologist. They also independently annotated approximately 10 cataract surgical videos and received feedback from the ophthalmologist before they started independent labelling. After initial annotation by this team, final annotation for all videos was adjudicated by a board-certified ophthalmologist to resolve discrepancies.

Individual frames were extracted from the video using OpenCV2[12] at a frame capture rate of 1 frame per second, yielding a total of 457,171 extracted frames. Frames were cropped to a square shape of resolution 256 × 256, centering the square over the original rectangular video frame. We split the dataset by video to prevent data leakage, reserving the frames from 211 videos for the training, 26 for the validation and 31 for the test set, with roughly 8:1:1 ratio for each surgeon. To ensure both validation and test set contains video with rare steps, we randomly assigned four videos containing advanced technique/other for each set. To represent the time information, the raw timestamp was extracted for each frame and normalized by dividing the maximum timestamp in each video. The normalization could mitigate the variation of the durations of different surgical phases between different levels of training. We also performed global mean centering by subtracting the mean pixel values in each channel across the whole training dataset.

Models 1: Transfer Learning from the VGG16 Architecture (VGG)[13]

We initialized the VGG16 network with weights pretrained on the ImageNet dataset.[14] We removed the original densely connected layers and output layer from the top of the VGG network and replaced it with two densely connected layers of 512 hidden nodes each, and a final softmax output layer that predicted which activity was represented in an individual video frame. The timestamp of the frame, normalized by the length of the video, was also used as an input to the model and was concatenated with the densely connected layer at the top of the model. We froze the first 15 layers of the VGG network up until Block 5 convolution 1, allowing parameters to be trained and updated for the remaining six layers (21,767,694 trainable parameters). The model architecture is summarized in Figure 1a.

Data augmentation was performed during the training process by transforming each frame with a random combination of rotation (≤180°), color channel shifting, brightening/dimming ± 20%, and zooming in/out ± 10%. Rare classes were upsampled during the training based on the following ratio: advanced techniques 4×, staining with trypan blue 8×, manipulation of iris 2×, injection into the subconjunctival space 2×, and 1× for the rest.

The model was trained with the Adam optimizer (learning rate 0.00001) and sparse cross entropy loss, for nine epochs, noting that overfitting occurred after that. The class with the highest output was defined

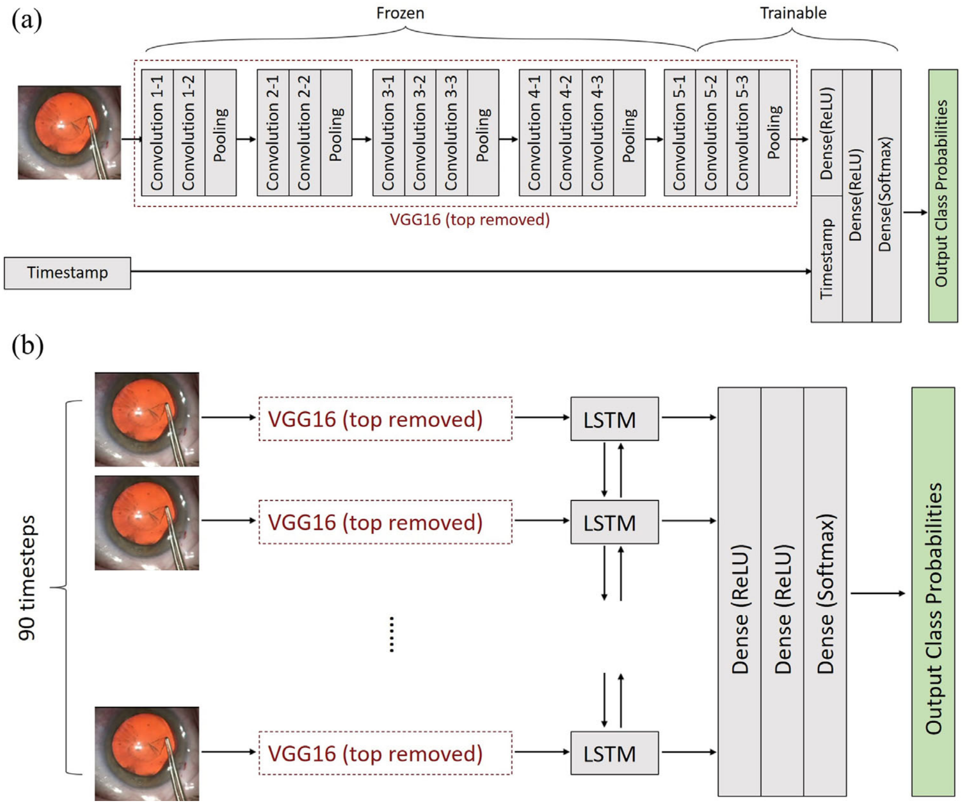


Figure 1. Deep learning model architectures for recognition of cataract surgical steps from surgical videos. Model architectures for (a) VGG and (b) CNN–RNN are shown. Top refers to the last three dense layers.

as the final prediction. Final predictions were further smoothed by averaging the predicted class probabilities across a continuous rolling window of five consecutive frames, to minimize the flicker of rapidly changing class predictions across neighboring frames.

Models 2: Convolutional Neural Network (CNN)–RNN

In the second model, we use the encoding from the last pooling layer of best VGG model as the input to our recurrent neural network (RNN), which is composed of bidirectional long short-term memory with 512 units.[15] Output from each long short-term memory cell was connected to two dense layers and one output layer with softmax activation. Adam optimizer was used with a learning rate fixed at 0.0001, beta1 0.9, and beta2 0.999. The loss function was sparse categorical cross entropy. The number of timesteps was tuned to 90, which means each input batch contained 90 consecutive frames, regardless of whether it included a phase transition or not. The model was trained until no decrease in validation loss in three consecutive epochs and was saved when it generated the lowest validation loss. Similar to VGG model, the output contained 14 softmax activation scores, and the class with the highest score was defined as the final prediction. The model architecture is summarized in Figure 1b. Both models were trained in Python 3.7.10 using tensorflow 2.4.1 and keras 2.4.3.

Evaluation Methods

Per-class receiver operating characteristic and precision recall (PR) curves along with areas under those curves and average precision scores were calculated. To summarize across the dataset, the same metrics were calculated but microaveraged across each class of activity. Microaveraging treats each observation equally regardless of which true class it belongs to and calculates the metric across the whole dataset, instead of first calculating the metric for each class and averaging them. This procedure puts more proportional weight on major classes and, therefore, represents a weighted average of the metric of interest. Frame-by-frame top-N accuracy was determined by examining whether the actual label was among the top N predicted classes with the highest predicted probabilities. We also evaluated per-class accuracy, sensitivity (recall), specificity, and precision (positive predictive value). Confusion matrices between true labels and model prediction were used for error analysis. Statistical analysis was performed using sklearn 0.22.1.

Class activation maps were visualized using Grad-CAM[16] by examining the last convolutional layer of the VGG network and the gradient information flowing into that layer to determine the critical regions to classify each class. A heatmap of highly activated pixels was overlaid with the original image for a random sample of correctly classified frames.

Results

The distribution of rare steps, advanced surgical techniques, and complications across the train, validation, and test videos are summarized in the Table, and the distribution of these between different surgeons are summarized in Supplementary Table S1. The number of videos contributed by five individual surgeons was 106, 97, 36, 18, and 2. A group of seven residents anonymously contributed nine additional videos. The overall classification accuracy was 76% for the VGG model and 84% for the CNN–RNN model. The VGG model had a weighted average precision of 77% and a weighted average sensitivity of 76%. CNN–RNN model had a weighted average precision of 85% and a weighted average sensitivity of 84%.

Using VGG, the area under receiver operating characteristic curves ranged from 0.82 in advanced

Table. Characteristics of Cataract Surgical Videos

No. of Videos Containing Advanced Cataract Surgery Steps	Total	Train	Validation	Test
No.	268	211	26	31
Anterior vitrectomy	5	3	1	1
Place capsule support	3	2	1	0
Trypan blue staining	111	95	5	11
Iris manipulation (malyugin ring, iris hooks, repositioning floppy iris)	41	32	4	5
Subconjunctival injection	62	50	4	8
Other/miscellaneous (limbal relaxing incisions, superior rectus traction suture, conjunctival vessel cautery, vitreous trimming at wound)	30	25	2	3

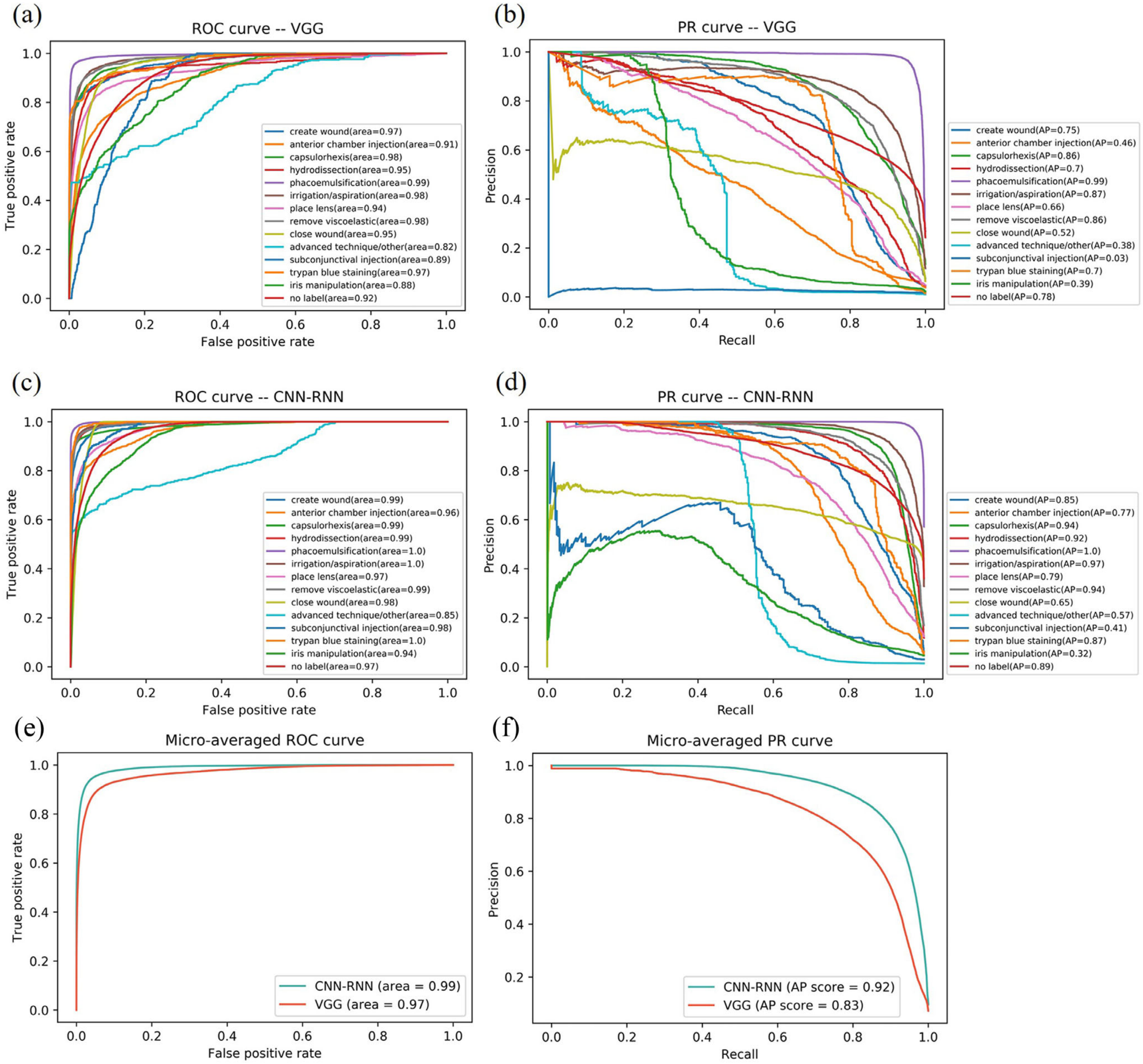
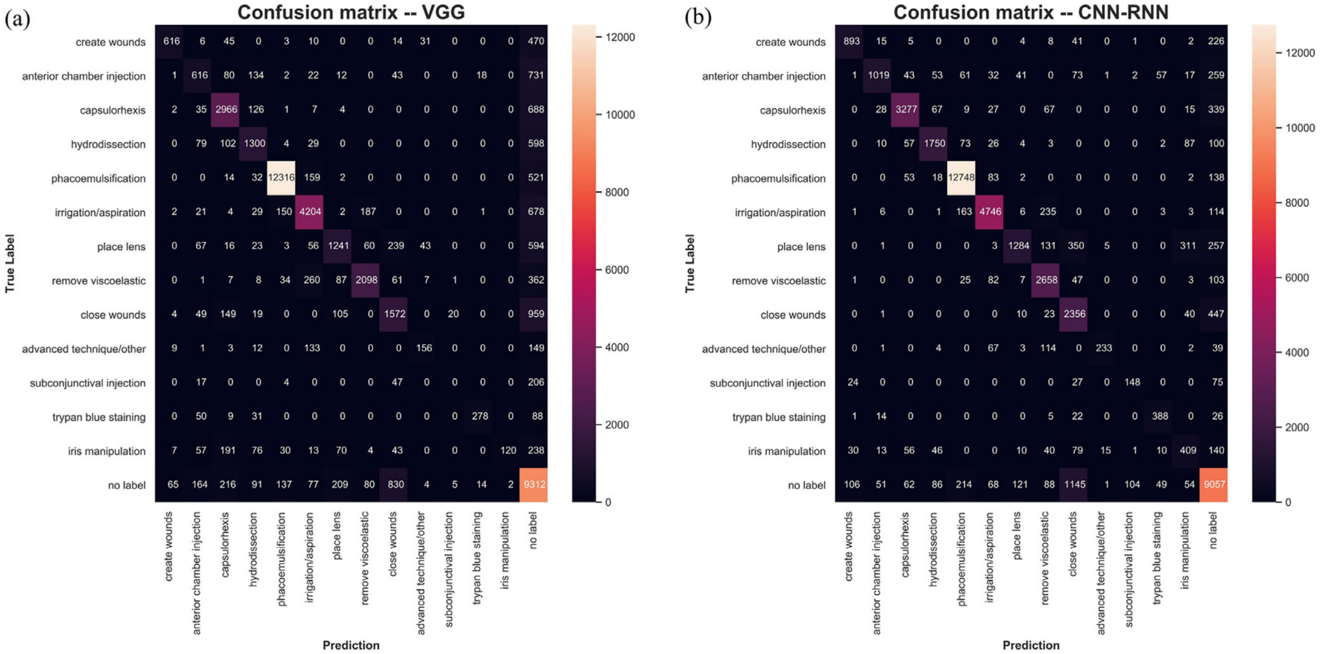


Figure 2. Receiver operating characteristic (ROC) and precision-recall (PR) curves for recognizing cataract surgical step. Per-class ROC curves with area under curves and PR curves with average precision (AP) scores for (a) (b) VGG and (c) (d) CNN–RNN model. (e) Microaveraged ROC curves and area under curves for both models. (f) Microaveraged PR curves and AP scores for both models.

technique/other to 0.99 in phacoemulsification, and that of CNN–RNN ranged from 0.85 in advanced technique/other to 1.00 in phacoemulsification, irrigation/aspiration, and trypan blue staining. The microaveraged area under receiver operating characteristic curve was 0.97 in VGG and 0.99 in CNN–RNN. The microaveraged average precision score was 0.83 for VGG and 0.92 for CNN–RNN. The receiver operating characteristic and PR curves are shown in Figure 2.

Per-class accuracy, sensitivity, specificity, and precision are summarized in Supplementary Table S2. Both models have the highest sensitivity in phacoemulsification, which were 0.944 and 0.977, respectively. Advanced technique/other, subconjunctival injection, and iris manipulation were the most difficult for the VGG model to predict, resulting in sensitivities of less than 50%. Notably, VGG did not predict any correct subconjunctival injection frames. The CNN–RNN model also struggled in these three classes, but the performance was improved. A simple (macro) average of step-specific accuracies of the VGG model was 96.5% and the CNN–RNN model 97.8%. Weighted



Confusion matrix -- VGG

True Label \ Prediction	create wounds	anterior chamber injection	capsulorhexis	hydrodissection	phacoemulsification	irrigation/aspiration	place lens	remove viscoelastic	close wounds	advanced technique/other	subconjunctival injection	trypan blue staining	iris manipulation	no label
create wounds	616	6	45	0	3	10	0	0	14	31	0	0	0	470
anterior chamber injection	1	616	80	134	2	22	12	0	43	0	0	18	0	731
capsulorhexis	2	35	2966	126	1	7	4	0	0	0	0	0	0	688
hydrodissection	0	79	102	1300	4	29	0	0	0	0	0	0	0	598
phacoemulsification	0	0	14	32	12316	159	2	0	0	0	0	0	0	521
irrigation/aspiration	2	21	4	29	150	4204	2	187	0	0	0	1	0	678
place lens	0	67	16	23	3	56	1241	60	239	43	0	0	0	594
remove viscoelastic	0	1	7	8	34	260	87	2098	61	7	1	0	0	362
close wounds	4	49	149	19	0	0	105	0	1572	0	20	0	0	959
advanced technique/other	9	1	3	12	0	133	0	0	0	156	0	0	0	149
subconjunctival injection	0	17	0	0	4	0	0	0	47	0	0	0	0	206
trypan blue staining	0	50	9	31	0	0	0	0	0	0	0	278	0	88
iris manipulation	7	57	191	76	30	13	70	4	43	0	0	0	120	238
no label	65	164	216	91	137	77	209	80	830	4	5	14	2	9312

Confusion matrix -- CNN-RNN

True Label \ Prediction	create wounds	anterior chamber injection	capsulorhexis	hydrodissection	phacoemulsification	irrigation/aspiration	place lens	remove viscoelastic	close wounds	advanced technique/other	subconjunctival injection	trypan blue staining	iris manipulation	no label
create wounds	893	15	5	0	0	0	4	8	41	0	1	0	2	226
anterior chamber injection	1	1019	43	53	61	32	41	0	73	1	2	57	17	259
capsulorhexis	0	28	3277	67	9	27	0	67	0	0	0	0	15	339
hydrodissection	0	10	57	1750	73	26	4	3	0	0	0	2	87	100
phacoemulsification	0	0	53	18	12748	83	2	0	0	0	0	0	2	138
irrigation/aspiration	1	6	0	1	163	4746	6	235	0	0	0	3	3	114
place lens	0	1	0	0	0	3	1284	131	350	5	0	0	311	257
remove viscoelastic	1	0	0	0	25	82	7	2658	47	0	0	0	3	103
close wounds	0	1	0	0	0	0	10	23	2356	0	0	0	40	447
advanced technique/other	0	1	0	4	0	67	3	114	0	233	0	0	2	39
subconjunctival injection	24	0	0	0	0	0	0	0	27	0	148	0	0	75
trypan blue staining	1	14	0	0	0	0	0	5	22	0	0	388	0	26
iris manipulation	30	13	56	46	0	0	10	40	79	15	1	10	409	140
no label	106	51	62	86	214	68	121	88	1145	1	104	49	54	9057

Figure 3. Confusion matrices for the deep learning models. Confusion matrices for (a) VGG and (b) CNN–RNN models.

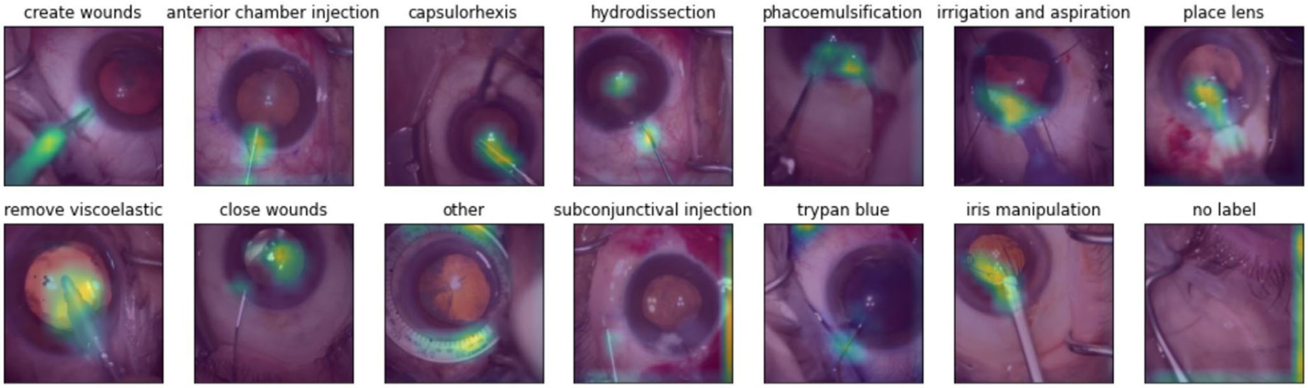


Figure 4. Gradient activation maps illustrating which pixels are most important to predict which cataract surgical step is being performed. A single frame from each surgical step was randomly sampled from the test set. A heatmap generated by Grad-CAM is overlaid on the frame, showing which pixels are highly activated in making the prediction for each class.

(micro) average of step-specific accuracies by label frequency of the VGG model was 93.9% and the CNN–RNN model 96.4%.

Confusion matrices for both models' predictions are shown in Figure 3. For most misclassifications, VGG most commonly predicted no label. The error rates were lower in the CNN–RNN model, but no label was still the most common misclassification. Top N accuracy with N ranging from 1 to 5 are illustrated in Supplementary Figure S1. Top 2 accuracies of CNN–RNN model improved to 94%, while VGG achieved a similar level at top 4 accuracies.

A Grad-CAM heatmap is shown in Figure 4. In classes with high accuracies, the model relied on the tool appearance, mostly tooltips, to predict surgical step. In contrast, the model did not look at the tool position in classes with low accuracies. For example, the model predicted close wound using information from the center of the pupil, which would not be expected to contain much information about close wound.

An example video timeline is shown in Figure 5, demonstrating the true timeline of events as well as the VGG and CNN–RNN predicted timelines. A companion video demonstrating activity recognition

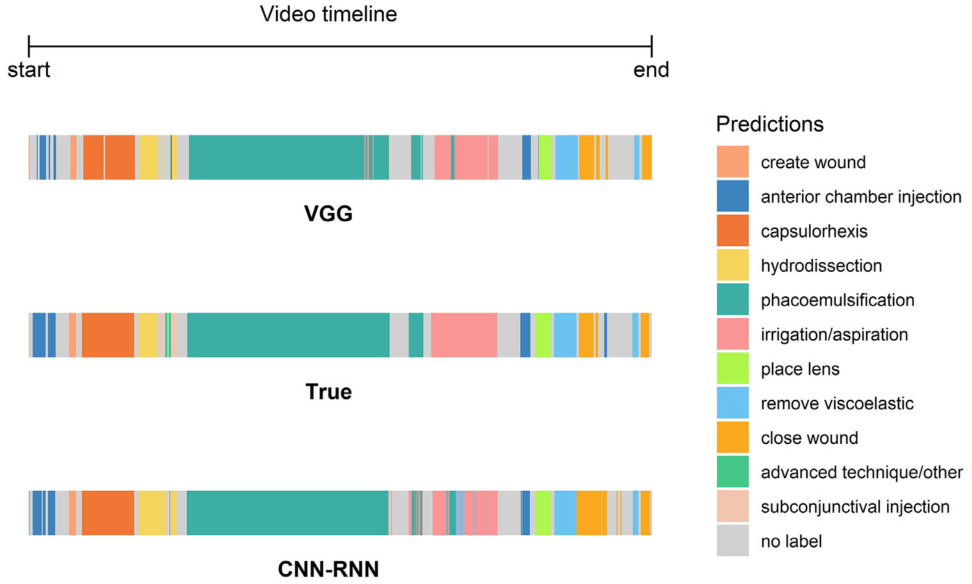


Figure 5. Video timeline comparison between ground-truth labels and models prediction. Model prediction and ground-truth labels for a randomly chosen video in the test set are shown as parallel timelines for comparison. CNN–RNN model produces more stable and accurate prediction than VGG model. (Note: The video recording started during the first paracentesis and captured only 1 second of it. As a result, for the true labels, the initial section for "create wound" is very slim and invisible.)

for an entire phacoemulsification movie is available as Supplementary Material.

Discussion

In this study, we developed, evaluated, and compared the performance of two deep learning models that identify the surgical step being performed in any given frame in a cataract surgery video. A novel aspect of our study was that we curated a larger and more varied corpus of videos with trainee surgeons from several different institutions. In addition, our videos were directly input into our models in an end-to-end manner, requiring minimal preprocessing, and they also included advanced techniques and complications, allowing us to predict the presence of these steps that had not been possible in previous studies. Our goal was to establish a more realistic estimation of the accuracy of deep learning models in this task. We also provided visualization of the gradient map to enhance interpretability and reliability of model prediction.

The VGG16 model with timestamp inputs generated reasonably good test accuracy of 76% and CNN–RNN model improved the overall test accuracy to 84%. The accuracy of our CNN–RNN outperformed previous work using hand-crafted features, which had achieved 72.9% and 81.2%.[17,18]

We found that a CNN alone could predict the core steps of cataract surgery well. Morita et al.[8] used Inception V3 to detect continuous curvilinear capsulorrhexis and nuclear extraction in real-time and achieved greater than 90% sensitivity. Our performance was higher in phacoemulsificationand other surgical steps, although performance was lower in capsulorrhexis. This finding could be attributed to the fact that our model has more steps to distinguish. For example, the similarity between cystotome and cannula used in anterior chamber injection could lead to misclassification.

A combination of RNN and CNN previously showed promising results in cataract surgical phase recognition as well.[9,10] Zisimopoulos et al.[9] trained ResNet-152 to predict the tools present in a frame and discovered that using encoding features after the last pooling layer to train RNN produced the best accuracy of 78%. Our approach is different in that we trained CNN directly to predict the phase instead of the tools present and achieved better overall accuracy. The best length of input sequences was found to be roughly 33 seconds in their study, which was less than 90 seconds, as in our study. This finding might indicate that inclusion of instrument labels provides richer information than plain video frame images for prediction of surgical step. Yu et al.[10] compared the performance of CNN and CNN plus RNN models using both video frame images and/or additional metadata, including labels indicating which surgical tools were present in the frames. The results showed that the RNN which incorporated the information about surgical tools, gave the highest accuracy. However, in deployment the model would then have to rely on preannotated information about surgical tools present in each frame, which would not be available generally. They also examined CNN and CNN plus RNN using only the unlabeled video frame images. Unlike our study, addition of the RNN to the architecture did not consistently improve the accuracies.[10] These studies only used datasets from a single institution and did not examine the model's performance on advanced surgical technique, which is essential to provide objective metrics in the presence of surgical complications or challenging case conditions. Our CNN–RNN reached significantly higher per-class accuracies, sensitivities, specificities, precisions, and areas under receiver operating characteristic curve than previous CNN plus RNN designs, possibly owing to a larger training dataset.

The frequent misclassification of no label as other surgical steps might be due to the subtle subjectiveness in determining the exact start time of an activity. Some actions involve instruments leaving the frame and reentering it, resulting in a small portion of interpolated frames that contain no instruments, which sometimes were mislabeled as other steps.

Another source of misclassification in the VGG model came from the rolling average procedure. Using five consecutive frames might over-smoothen short surgical steps that contained fewer than five frames. We noticed 2% and 3% drops of sensitivity for the typically shorter steps of create wound and anterior chamber injection, respectively. Nonetheless, this rolling average improved sensitivities by 1% to 7% for longer steps, thus making the overall sensitivity higher. Of note, the rolling average procedure was only used for the VGG model and was not required for the CNN–RNN model which automatically provides some smoothing for predictions of temporally adjacent frames.

The model's reliance on seeing the instruments present in the frame to produce its prediction is reassuring, because humans would also use the presence of certain instruments as a clue to determine which surgical step is occurring. However, one possible limitation of this behavior is that the ability of the model to differentiate between surgical steps could be constrained by instruments with similar appearances. For example, the tiny visual difference between the cystotome and the injection cannula, sometimes only a matter of seeing a few extra pixels at the end of the cystotome, might be one of the reasons why the model could not differentiate well between capsulorrhexis and anterior chamber injection. Theoretically, higher resolution images could help to solve this problem by simply presenting more pixels to the model, but they also contribute to higher computational burden. This behavior also induces the risk of misclassification with different brands of instruments in different countries, regions, or surgeons and may miss useful information in other parts of the image that human surgeons would also rely upon, such as the presence of a fragmented lens or an intraocular lens.

Our study has high translational relevance for the improvement of surgical training and development of automated surgical systems. For example, a simple but valuable application would be to automatically calculate the time a surgeon spends in each step, which is correlated with a surgeon's performance.[19] After each surgery, residents could easily use our model to track surgical times on a granular level over time and determine which steps he or she might need further improvement. In addition, the ability to recognize surgical steps is a key step toward the development and implementation of context-aware computer-assisted surgery[20,21] or even fully robotic surgery. Such systems can be used to perform context-aware adaptation of device settings, provide danger notifications to surgeons, and deliver real-time insights. Another future application is applying a similar strategy to train models to recognize the steps of other ophthalmic surgeries, such as glaucoma surgeries.

Our model directly learned useful features from images during the training process, which has become one of the most common data types used in surgical phase recognition in recent years.[22] This avoids the need to extract other information, such as instrument use, color histograms, texture, etc. before running the model and allows faster and easier model deployment.

Our study has several limitations. First, because we included complete cataract surgical videos in the training data, in future applications residents would have to input the whole-length video to achieve the same model performance. Second, the relatively few videos with rare surgical steps (such as anterior vitrectomy, placement of capsular tension ring, etc.) caused significant difficulty for model prediction despite aggressive upsampling of these video frames. Amassing more examples of these types of actions would be ideal to further improve performance on these steps before model deployment. Third, the relatively long time sequence needed for our CNN–RNN model currently excludes the possibility of deployment in a real-time prediction setting. Fourth, the level of training of each surgeon in our dataset is uncertain because of the semianonymous collection of videos, which may limit the robustness of the model. Last, our models only take static images as inputs and do not take advantage of the valuable information of object motion from videos. Incorporating movement-based models such as two-stream convolutional networks[23] could have potential for performance improvement.

In summary, by using a larger and more varied dataset, our deep learning model with a CNN plus RNN architecture showed highly accurate predictions for routine steps of cataract surgery and gives realistic estimates on how the model might perform in diverse cataract surgeries with advanced surgical steps.

Acknowledgments

Supported by NLM T15 LM 007033, Career Development Award from Research to Prevent Blindness, unrestricted departmental grants from Research to Prevent Blindness, the National Eye Institute P30-EY026877. This project utilized credits for Google Cloud Platform for computing. The funders had no role in the study design, conduct, or decision to publish.

Disclosure: **H.-H. Yeh**, None; **A.M. Jain**, None; **O. Fox**, None; **S.Y. Wang**, None

References

1. Pascolini D, Mariotti SP. Global estimates of visual impairment: 2010. *Br J Ophthalmol.* 2012;96(5):614–618.
2. Cullen KA, Hall MJ, Golosinskiy A. Ambulatory surgery in the United States, 2006. *Natl Health Stat Report.* 2009(11):1–25.
3. Dooley IJ, O'Brien PD. Subjective difficulty of each stage of phacoemulsification cataract surgery performed by basic surgical trainees. *J Cataract Refract Surg.* 2006;32(4):604–608.
4. Gensheimer WG, Soh JM, Khalifa YM. Objective resident cataract surgery assessments. *Ophthalmology.* 2013;120(2):432–433.e431.
5. Smith P, Tang L, Balntas V, et al. "PhacoTracking": an evolving paradigm in ophthalmic surgical training. *JAMA Ophthalmol.* 2013;131(5):659–661.
6. Yeung S, Rinaldo F, Jopling J, et al. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *NPJ Digit Med.* 2019;2:11.
7. Jin A, Yeung S, Jopling J, et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. Paper presented at: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); Lake Tahoe, NV:USA, March 12–15, 2018.
8. Morita S, Tabuchi H, Masumoto H, Yamauchi T, Kamiura N. Real-time extraction of important surgical phases in cataract surgery videos. *Sci Rep.* 2019;9(1):16590.
9. Zisimopoulos O, Flouty E, Luengo I, et al. *DeepPhase: Surgical Phase Recognition in CATARACTS Videos*. Cham, September 16, 2018.
10. Yu F, Silva Croso G, Kim TS, et al. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA Netw Open.* 2019;2(4):e191860.
11. Dutta A, Zisserman A. The VIA annotation software for images, audio and video. Proceedings of the 27th ACM International Conference on Multimedia; October 21–25, 2019; Nice, France.
12. Bradski G. The OpenCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer.* 2000;25(11):120–123.
13. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014; arXiv:1409.1556, https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S. Accessed September 01, 2014.
14. Deng J, Dong W, Socher R, Li L, Kai L, Li F-F. ImageNet: a large-scale hierarchical image database. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition; Miami, FL, June 20–25, 2009.
15. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput.* 1997;9(8):1735–1780.
16. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Grad-CAM Batra D.: Visual explanations from deep networks via gradient-based localization. Paper presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Venice, Italy, October 22–29, 2017.
17. Quellec G, Charriere K, Lamard M, et al. Real-time recognition of surgical tasks in eye surgery videos. *Med Image Anal.* 2014;18(3):579–590.
18. Quellec G, Lamard M, Cochener B, Cazuguel G. Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials. *IEEE Trans Med Imaging.* 2015;34(4):877–887.
19. Schoeffmann K, Taschwer M, Sarny S, Münzer B, Primus MJ, Putzgruber D. Cataract-101: video dataset of 101 cataract surgeries. Proceedings of the 9th ACM Multimedia Systems Conference; Amsterdam, the Netherlands; June 12–15, 2018.
20. Franke S, Rockstroh M, Hofer M, Neumuth T. The intelligent OR: design and validation of

a context-aware surgical working environment. *Int J Comput Assist Radiol Surg*. 2018;13(8): 1301–1308.

21. Kitaguchi D, Takeshita N, Matsuzaki H, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surg Endosc*. 2020;34(11):4924–4931.
22. Garrow CR, Kowalewski K-F, Li L, et al. Machine learning for surgical phase recognition: a systematic review. *Ann Surg*. 2021;273(4):684–693.
23. Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. 2014; arXiv:1406.2199.