AMERICAN ACADEMY OF OPHTHALMOLOGY®

CatStep: Automated Cataract Surgical Phase Classification and Boundary Segmentation Leveraging Inflated 3D-Convolutional Neural Network Architectures and BigCat

Ossama Mahmoud, BSc,[1,2] *Han Zhang, BSc,*[3] *Nicholas Matton, MS,*[4] *Shahzad I. Mian, MD,*[1] *Bradford Tannen, MD, JD, MBA,*[1] *Nambi Nallasamy, MD*[1,5]

Objective: Accurate identification of surgical phases during cataract surgery is essential for improving surgical feedback and performance analysis. Time spent in each surgical phase is an indicator of performance, and segmenting out specific phases for further analysis can simplify providing both qualitative and quantitative feedback on surgical maneuvers.

Study Design: Retrospective surgical video analysis.

Subjects: One hundred ninety cataract surgical videos from the BigCat dataset (comprising nearly 4 million frames, each labeled with 1 of 11 nonoverlapping surgical phases).

Methods: Four machine learning architectures were developed for segmentation of surgical phases. Models were trained using cataract surgical videos from the BigCat dataset.

Main Outcome Measures: Models were evaluated using metrics applied to frame-by-frame output and, uniquely in this work, metrics applied to phase output.

Results: The final model, CatStep, a combination of a temporally sensitive model (Inflated 3D Densenet) and a spatially sensitive model (Densenet169), achieved an F1-score of 0.91 and area under the receiver operating characteristic curve of 0.95. Phase-level metrics showed considerable boundary segmentation performance with a median absolute error of phase start and end time of just 0.3 seconds and 0.1 seconds, respectively, a segmental F1-score @70 of 0.94, an oversegmentation score of 0.89, and a segmental edit score of 0.92.

Conclusion: This study demonstrates the feasibility of high-performance automated surgical phase identification for cataract surgery and highlights the potential for improved surgical feedback and performance analysis.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2024;4:100405*

Supplemental material available at www.aaojournal.org.

Cataract surgery is an essential part of ophthalmic surgical training and practice. Although advances in cataract surgery technology such as foldable intraocular lenses and active fluidics have led to decrease in complication rates, existing methods of providing feedback to trainee surgeons on surgical performance remain limited.

Because patients are typically awake for cataract surgery, providing verbal intraoperative feedback to trainee surgeons is problematic. Providing postoperative feedback is challenging due to the need to move between cases efficiently. Furthermore, verbal feedback from faculty is usually qualitative in nature, may lack objectivity, and often lacks a longitudinal perspective on trainee progress. Accordingly, simplifying and automating approaches to providing objective feedback on trainee surgical performance is likely to be of value.

Identifying surgical phase from video recordings offers 4 key potential applications in improving surgical feedback and performance analysis. First, it offers the ability to quantify the time spent in each surgical phase, which itself can be an important indicator of how well a surgeon is performing.[1] Second, the ability to segment out specific phases for further analysis can simplify the process of providing both qualitative and quantitative analyses of specific surgical maneuvers. For example, a trainee struggling with performing the capsulorrhexis, a delicate and challenging step in cataract surgery, could easily obtain a "supercut" of each capsulorrhexis they have performed for their own review as well as review by their surgical mentors. Third, surgical phase identification is a fundamental building block for providing more complex automated surgical feedback for an individual phase of surgery across a large

https://doi.org/10.1016/j.xops.2023.100405
ISSN 2666-9145/23

set of surgeries. To provide automated longitudinal feedback on the fluidity of anterior capsulotomy creation to a resident, for example, identification of the start and end times of the capsulorrhexis phase must be performed for hundreds of surgical recordings. Fourth, surgical phase identification is essential for the tracking of general metrics of surgical skill (e.g., eye centration and operating microscope focus) to assess their changes during various steps of surgery. Accordingly, we sought to develop and validate a system for automating the segmentation of raw ophthalmic surgical video into its component steps or phases.

Although the use of machine learning (ML) models to identify surgical phases has been attempted in various forms, each approach has had limitations that would prevent their application for the purposes described above. The primary limitation among these prior efforts has been the lack of automated detection of phase start and end times, in favor of focusing on the more straightforward task of frame-level classifications of surgical phase. The true segmentation of raw video into component steps with a system that outputs phase start and end times is a necessary step toward automating surgical phase-level analyses. Noise in frame-level classifications makes the task of outputting phase start and end times substantially harder than outputting frame-level classifications alone.

A prior study by Yu et al,[2] used human-defined phase start and end times to analyze phase classification performance of 5 ML models. Using a dataset of 100 cataract surgery videos, the classifiers were trained to perform a single-class classification among 10 surgical phases at the frame level, with performance ranging from area under the curve of 0.712 to 0.773 using 5-fold cross-validation. The use of human-defined phase start and end times precluded the calculation of phase-level segmentation performance metrics.

In another study, Zisimopoulos et al[3] trained a residual neural network followed by a recurrent neural network (RNN) to identify surgical phases with a training set of 25 cataract surgery videos, achieving a maximum accuracy of 78% in frame-level classification. In a more recent study, Garcia Nespolo et al[4] trained a convolutional neural network (CNN)-based model on 6 surgical videos to identify 3 surgical phases as part of a larger resident feedback system. Both studies were limited by small datasets and the lack of phase-level start and end time predictions.

Recently, work has also been reported on surgical phase identification in nonophthalmic surgery. In a study by Sahu et al[5] from 2020, 80 laparoscopic surgery videos were used to develop a 7-class surgical phase classifier. Their approach utilized surgical tool data passed into a long short-term memory-based architecture (ZIBNet), though again was limited to frame-level predictions. In another study focusing on laparoscopic surgery, Zhang et al[6] created a 3-dimensional convolutional neural network (3D-CNN) combined with a sequence-to-sequence model capable of calculating timestamp-based predictions for 5 phases. However, frame-level performance was limited, with a maximum F1-score of 0.74 across all architectures studied. This translated into limited phase-level segmentation performance, with an event ratio (closer to 1 indicating better performance) of 0.342.

In the present work, we describe the development and validation of a system for automating the segmentation of raw cataract surgery video into component phases. In this work, we have attempted to overcome limitations of prior efforts in 3 primary ways. First, we developed the largest cataract surgery phase annotation dataset (BigCat), containing nearly 4 million frames, reported to date. Second, we attempted to improve classification performance through the use of ML architectures capable of modeling both spatial and temporal relationships in the data. In particular, we investigated if inflation of 2-dimensional convolutional networks into 3 dimensions would allow for the learning of spatiotemporal feature extractors from video while leveraging architectures and parameters successful in our prior work on cataract surgical instrument identification.[7] Third, we sought to more thoroughly analyze the phase-level segmentation performance of our models and provide a sense of the direct applicability of these models to the task of fully automated surgical phase segmentation from raw surgical video.

Materials and Methods

Data Collection

Cataract surgical videos were collected at the University of Michigan Kellogg Eye Center between 2020 and 2021. Institutional review board approval was obtained for the study (HUM00160950), and it was determined that informed consent was not required because of its retrospective nature and the anonymized data utilized in this study. The study was carried out in accordance with the tenets of the Declaration of Helsinki. The BigCat database was developed from the surgical videos gathered and was described in detail previously.[7] The BigCat database consists of a fully annotated set of cataract surgeries performed by attending surgeons at the University of Michigan Kellogg Eye Center. For this study, femtosecond laser cataract surgeries and complex cataract surgeries (those qualifying for Current Procedural Terminology code 66982) were excluded so as to ensure a standardized set of surgical phases. Cases with incomplete recordings were also excluded. Segments from before surgery and after surgery were trimmed, but video during surgery was otherwise completely unedited. The source resolution was 1920 × 1080 pixels at a frame rate of 30 frames per second. A total of 208 videos were selected for annotation of surgical phase ground truth (GT) for every frame. Eleven distinct nonoverlapping surgical phases (listed in Table 1) were annotated with a binary designation for each phase for each frame. Phase annotations were performed (after training by NN) by a third-party annotation services provider (Alegion Inc). All phase annotations were validated manually by the research team prior to inclusion in the dataset. One hundred ninety videos passed annotation validation checks to ensure appropriate and complete annotations for all available frames. Table S2 provides a comparison of BigCat with other reported cataract surgery video datasets.[8–14]

Data Processing

Videos were resized to 3 different resolutions (135 × 68, 240 × 135, and 480 × 270 pixels) to explore the tradeoffs among video

Table 1. Phases and Their Durations Within the BigCat Dataset

Phase	Average Time Spent in Phase, s	Average Phase Length as Percentage of Total Video
No Activity	74	11%
Paracentesis	11	2%
General Injection	51	8%
Main Wound	16	2%
Capsulorrhexis Initiation	25	4%
Capsulorrhexis Formation	38	5%
Hydrodissection	40	6%
Phacoemulsification	244	35%
Cortical Removal	88	13%
Lens Insertion	29	4%
Viscoelastic Removal	49	7%
Wound Closure	27	4%
Total	**692**	**100%**

Bolded values show the average total length of a surgical video.

size, model training and inference time, and model phase recognition performance. To improve the generalizability of the models studied, input data were augmented by applying random transformations, including rotations, shifts, shears, zooms, horizontal flips, and rescales. Of the 190 videos that passed validation checks, 114 videos (2 282 382 frames) were allocated for training, 38 videos (838 005 frames) were allocated for validation, and 38 videos (826 266 frames) were held out for testing.

To construct the sequences of frames, or clips, to be used as input to the 3D-CNN based inflated 3-dimensional (I3D) model, a 30-frame look-back window was taken before each frame, and the stride of the look-back window was seen as a hyperparameter and tuned.

Model Development

As mentioned above, the problem of phase identification was defined at both the frame and phase level. Frame-level identification was defined as the task of predicting the phase for a given frame of surgical video utilizing only that frame or a sequence of frames leading up to and including the frame in question. This task could be considered a multi-class single-label classification task with 12 possible classes, 11 active phases, and a "No Activity" phase indicating the absence of active surgical activity. Phase-level segmentation was defined as the task of identifying start and end times for each of the 11 active cataract surgery phases listed in Table 1 for a given complete raw surgical video.

The primary hypothesis guiding our approach to model development for the aforementioned tasks was that both spatial and temporal features are relevant for phase identification. This hypothesis is founded on the knowledge that certain cataract surgery instruments are present in multiple phases, including cannulas (injections, hydrodissection, and wound closure) and the irrigation-aspiration handpiece (cortical removal and viscoelastic removal). Accordingly, it would be expected that instrument trajectories through the surgical field would be valuable in distinguishing phases with similar frame-level spatial characteristics. To test this hypothesis, we developed a total of 4 model architectures for frame-level identification. For phase-level segmentation, all architectures leveraged the "No Activity" label as a way to determine phase start and end times.

The first algorithm considered consisted of a CNN, a dense neural network (NN), and a softmax function. The CNN was used to draw spatial patterns from the input images, whereas the dense NNs were meant to make predictions on the input images.

The softmax mapped these predictions into a probability between 0 and 1. The output was a vector indicating the probabilities that each of the 11 surgical phases were represented by the input frame. The CNN used was the Densenet169 because of its densely connected network, which mitigates the vanishing gradient problem and promotes feature reuse.[5] Weights from ImageNet pretraining were used to initialize the Densenet169 model. To incorporate time dependencies and address smoothness of the CNN predictions, our second approach involved the addition of a recurrent neural network (RNN) on top of our CNN model. The RNN implemented consisted of a fully connected Simple-RNN followed by a dense NN layer and a softmax function.

To incorporate both temporal and spatial information into a single model, a 3D-CNN was considered as our third architecture. The 3D-CNN architecture was based on the I3D model.[15] The I3D approach applies an inflation process to a deep CNN allowing for the learning of spatiotemporal feature extractors. In this process, each individual filter in the CNN is inflated by duplicating the weights of the filter and stacking them as a 3-dimensional filter. We applied this inflation process to our pretrained phase detection Densenet model to generate a 3D-CNN capable of learning directly from video clips (Fig. 1). Data blocks were then generated from the video frame data, with the third dimension of the block acting as the time dimension. A single label was assigned to the block using the value of the last frame in the 3D block. Frame rate, image size, and number of frames per block were varied and tuned.

The fourth model considered was an ensemble combining the Densenet and I3D models. The data pipeline for the ensemble model is depicted in Figure 2. The input to the Densenet model was a single frame of resolution 240 × 135, whereas the input to the I3D model was a 30-frame clip of the same resolution, ending with the same frame as the input to the Densenet model. This design ensured that the frame label was consistent across both inputs, enabling the consolidation of the output from each model into a single ensemble model. The ensemble was constructed by passing the output of the second to last layer of each model (vectors of size 256) and a normalized frame number through a fully connected NN. The fully connected network consisted of 2 layers with 128 and 32 nodes each. A grid search was performed to optimize learning rate and clip frame rate. Learning rates of 1e-4, 5e-4, 1e-3, and 5e-3 and frame rates of 30, 15, 10, 5, 3, and 1 frame per second were tested.

The trade-off between various model architectures' performance on the validation set and model complexity was quantified. To speed up model development, only 25% of our training set (still representing over 500 000 frames) was utilized during model selection. Once the optimal model and hyperparameters were found, we trained our final model with all available training data.

Phase Start and End Times

Because noise in frame-level phase predictions could lead to fragmentation of phase-level predictions, a mean filter was applied to the predictions generated by each ML model for phase-level predictions. A sliding window of 30 frames was utilized. All phases were bounded between regions with "No Activity," allowing for segmentation of each phase and determination of start and end times. Start and end times for each of the 11 active phases were determined by first identifying and segmenting contiguous "Activity" phases. To accomplish this, the frame-level predictions of "No Activity" from each model were utilized. Every set of contiguous frames with a "No Activity" label were considered as the boundaries for the "Activity" phases. The regions containing activity were then assigned a label corresponding to the most common label between the activity's start

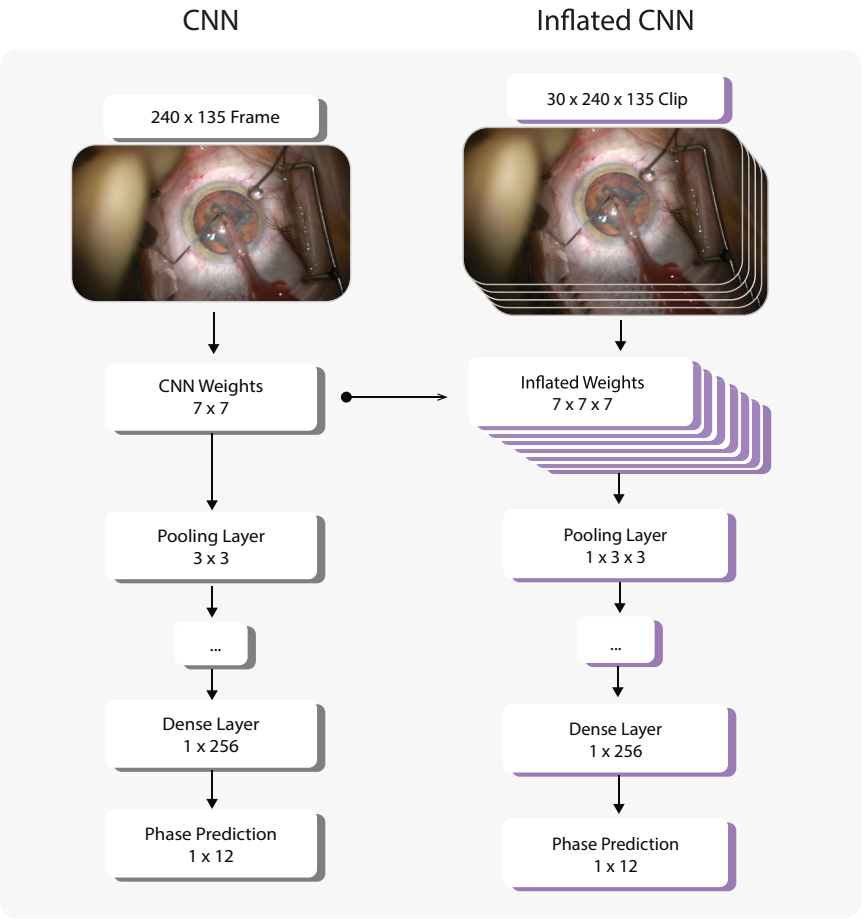

Figure 1. Densenet inflation process, showing convolution layers expanded by adding an additional symmetric dimension (e.g., 7×7 to $7 \times 7 \times 7$), that gets filled by duplicating the weights across the additional dimension. Pooling layers are unaffected by the inflation operation (e.g., 3×3 to $1 \times 3 \times 3$).

and end time, as identified by the phase prediction model. If a "No Activity" phase separated 2 activity phases with the same predicted label, the 2 phases were treated as a contiguous phase and collapsed into a single phase.

Model Evaluation

Model performance was evaluated using a wide range of metrics. Frame-level prediction metrics included class accuracy, recall, precision, F1-score, and area under the receiver operating characteristic curve (AUROC). Class-level recall refers to the proportion of instances within a specific surgical phase that the model correctly classifies among all instances of that phase, providing insight into the model's ability to capture true positives for that phase. Precision, on the other hand, denotes the fraction of instances classified as a particular surgical phase that are actually accurate, aiding in understanding the model's capacity to minimize false positives for that phase. The F1-score combines both precision and recall, offering a single metric that balances these 2 aspects and provides an overall assessment of the model's frame-level performance in identifying the surgical phases.

To assess a model's phase-level segmentation performance from a variety of different perspectives, 5 different metrics were considered. The first 2 were the mean absolute error (MAE) and median absolute error in the predicted start/end time of each phase compared to the true start/end time of each phase. These metrics provide a sense of the offset between predicted and GT phase boundaries and are most relevant when a high level of accuracy in human phase annotation is expected, as with BigCat.

The third phase-level metric considered was the Over-Segmentation Score (S_O). The Over-Segmentation Score measures the extent of overlap between GT and predicted segments.[16] This score is a function of the predicted segment with a maximum intersection over union for a given GT segment and is given by:

$$S_O(G,P) = \frac{1}{N}\sum_{i=0}^{N} \max \left| \frac{G_i \cap P_j}{G_i \cup P_j} \right|$$

where $G = \{G_0 \ldots G_i \ldots G_N\}$ is the sequence of GT phases, and $P = \{P_0 \ldots P_j \ldots P_N\}$is the set of phase predictions. The Over-Segmentation Score lies within [0, 1] and a higher value indicates better performance. As the name indicates, this score penalizes over-segmentation errors, in which there are multiple predicted segments within 1 true segment.

The fourth phase-level metric considered was the segmental F1-score (F1@k). The F1@k metric is calculated by first computing the intersection over union for each predicted phase with respect to its corresponding GT phase and subsequently using the given intersection over union threshold, k, to determine whether each predicted phase is considered a true positive or a false positive.[17] Those GT phases without a corresponding prediction are considered false negatives. The precision and recall with threshold k are then computed by summing true positives, false

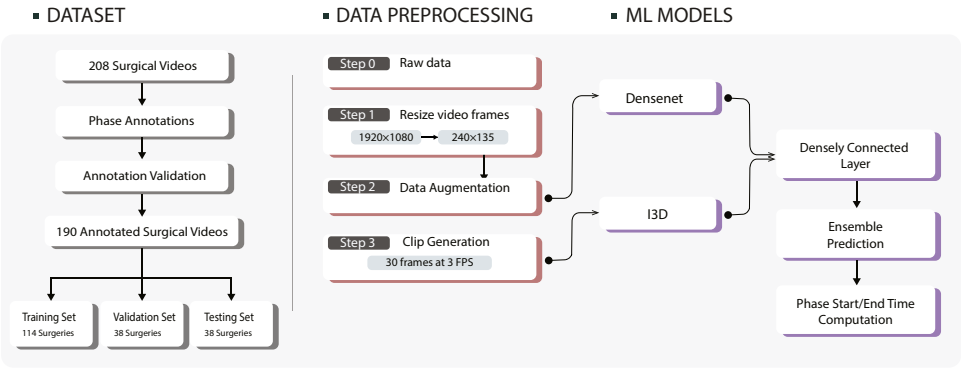

Figure 2. Data pipeline for the ensemble model. FPS = frames per second; I3D = inflated 3-dimensional; ML = machine learning.

positives, and false negatives across all classes. The F1@k is then computed in the traditional manner from the summed precision and recall determined using threshold k:

$$P@k = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} TP_c + \sum_{c=1}^{C} FP_c}$$

$$R@k = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} TP_c + \sum_{c=1}^{C} FN_c}$$

$$F1@k = \frac{2 * P@k * R@k}{P@k + R@k}$$

The magnitude of the reduction in F1@k as k increases indicates the degree of overlap between the predicted phases and the GT phases. Smaller reductions in F1@k as k grows indicate greater overlap.

The fifth phase-level metric considered was the segmental edit score (SES). The SES measures how well a model predicts the ordering of phases independent of slight time offsets.[16,18] Specifically, the SES allows for the evaluation of misclassifications, insertions, and deletions in phase predictions. To compute the SES, an edit distance is first calculated by identifying the minimum number of substitutions, deletions, and/or insertions required to transform the sequence of predicted phases (P) into the sequence of GT phases (G) using the Wagner–Fischer algorithm. This edit distance is then normalized by the greater sequence length among P and G. That is:

$$SES = 1 - s_e(G,P) = 1 - \frac{EditDistance(G,P)}{\max(|G|,|P|)}$$

As with the Over-Segmentation Score and F1@k, the SES lies on the interval [0, 1], and a higher score indicates better performance.

Table 3. Validation Metrics Comparing Performance of the Densenet Model With Varying Input Resolution

Input Resolution	F1-Score	Accuracy	AUROC	Precision	Recall
120 × 68	0.85	0.98	0.91	0.86	0.83
240 × 135	0.90	0.99	0.94	0.91	0.91
480 × 270	0.91	0.99	0.94	0.91	0.89

AUROC = area under the receiver operating characteristic curve.

Statistical Analysis

Differences in model performance on the validation set were assessed using the Friedman test, followed by post hoc paired Wilcoxon signed-rank tests with Bonferroni correction.

Implementation

Data pipelines and ML models were developed and tested in Python 3.8 with TensorFlow 2.2.0 and Keras 2.3.0. Testing, including inference time measurements, was performed using a machine with 4 Nvidia RTX 2080 Ti Graphics Processing Units. For each test run, we utilized 2 Graphics Processing Units to load the model, load the testing data, and make inferences on the testing data.

Results

Dataset Characteristics

A final dataset consisting of annotated video recordings of 190 cataract surgeries performed at University of Michigan's Kellogg Eye Center was gathered. The source resolution was 1920 × 1080 pixels at a frame rate of 30 frames per second with a mean duration of 692 seconds and standard deviation of 161 seconds. As seen in Table 1, the longest phase was phacoemulsification, taking 244 seconds or 35% of each surgery on average. The shortest phase was paracentesis creation, taking approximately 11 seconds or just 2% of the total length of the procedure on average.

Table 4. Validation Metrics for the Four Models Considered

Model	F1-Score	Accuracy	AUROC	Precision	Recall
Densenet	0.90	0.99	0.94	0.911	0.89
RNN-densenet	0.88	0.99	0.94	0.85	0.85
I3D	0.91	0.99	0.94	0.91	0.91
I3D-densenet ensemble	0.91	0.99	0.95	0.91	0.91

AUROC = area under the receiver operating characteristic curve; I3D = inflated 3-dimensional; RNN = recurrent neural network.
Results were obtained using an input resolution of 240 × 135 pixels. Phase start and end time errors were averaged across all activity phases considered.

Table 5. Class-wise and Overall Validation F1-Scores for the Models Studied

Phase	Densenet	RNN-Densenet	I3D	I3D-Densenet Ensemble
No activity	**0.83**	0.76	0.81	**0.83**
Paracentesis	0.86	0.81	0.89	**0.88**
General injection	**0.85**	0.76	0.82	0.84
Main wound	**0.91**	0.86	0.89	**0.91**
Capsulorrhexis initiation	0.84	0.81	0.87	**0.90**
Capsulorrhexis completion	0.94	0.93	0.94	**0.95**
Hydrodissection	**0.93**	0.90	0.89	0.91
Phacoemulsification	**0.99**	0.98	**0.99**	**0.99**
Cortical removal	0.95	**0.97**	**0.97**	**0.97**
Lens insertion	**0.91**	0.90	0.88	0.90
Viscoelastic removal	**0.96**	0.92	0.95	**0.96**
Wound closure	0.87	0.87	0.89	**0.91**
Average	0.90	0.88	0.91	**0.91**

I3D = inflated 3-dimensional; RNN = recurrent neural network.
The highest-performing model or models for each phase are in bold.

Model Performance

Varying image size appeared to have an impact on the Densenet model's performance, with validation F1-scores rising from 0.85 to 0.91 when increasing input image resolution from 120 × 68 pixels to 480 × 270 pixels. Although increasing input image resolution yielded higher performance, there were diminishing returns (see Table 3). As such, the 240 × 135 pixel resolution was selected as the standard resolution across all models for further testing to balance space constraints with performance.

All 4 of the models described in the model development section demonstrated considerable frame-level phase classification F1-scores (0.88−0.91), accuracies (0.99−0.99), and AUROC (0.94−0.95) on the validation set, as seen in Table 4. However, the Densenet slightly outperformed both the I3D model and the RNN-Densenet model overall. The various models performed differently in identifying particular phases of surgery. Although the Densenet outperformed the I3D model and RNN-Densenet on several phases, the I3D model outperformed the Densenet and RNN-Densenet in identifying capsulorrhexis initiation by 7%.

Class-wise frame-level classification performance is summarized in Table 5.

The I3D-Densenet Ensemble model outperformed the individual models (Table 4) and as such was chosen for further development. The difference in performance between the Ensemble and each of the other models was statistically significant (see Table S6).

The I3D-Densenet Ensemble model exhibited a longer training time of 0.054 seconds per instance, in comparison to the Densenet (0.015 seconds per instance), Densenet-RNN (0.015 seconds per instance), and I3D model (0.036 seconds per instance). Additionally, it demonstrated longer inference times. The detailed training and inference times can be found in Table S7.

After hyperparameter tuning, the optimized I3D-Densenet Ensemble model was trained on both the training and validation sets and termed CatStep. The CatStep model was evaluated on a hold-out testing set, the results of which are presented in Table 8. The CatStep model achieved an F1-score of 0.91, accuracy of 0.99, and an AUROC of 0.95. The final model demonstrated strong phase-level segmentation performance as well, with a start time MAE of 2.3 seconds and end time MAE of 1.6 seconds. Figure 3 depicts a representative surgical video's predicted phase segmentation compared with GT. The CatStep model achieved an Over-Segmentation Score of 0.89, F1@50 of 0.95, and SES of 0.92.

Discussion

Automated segmentation of surgical phases is a crucial step in the delivery of automated analysis of surgical performance at scale. Systems seeking to provide surgical feedback to trainees or analyze performance for quality assurance will require the ability to identify specific surgical maneuvers for downstream analysis. The CatStep model

Table 8. Final Model (Ensemble Model With Densenet169 Combined With I3D Model) Performance on Hold-Out Test Set

Metric	Final Model (CatStep) Performance
Frame-Level Metrics	
F1-score	0.91
Accuracy	0.99
AUROC	0.95
Precision	0.90
Recall	0.91
Phase-Level Metrics	
Start Time MAE, MedAE (s)	2.3, 0.3
End Time MAE, MedAE (s)	1.6, 0.1
Over-segmentation score	0.89
F1@k	
F1@50	0.95
F1@60	0.95
F1@70	0.94
Segmental edit score	0.92

AUROC = area under the receiver operating characteristic curve; MAE = mean absolute error; MedAE = median absolute error.

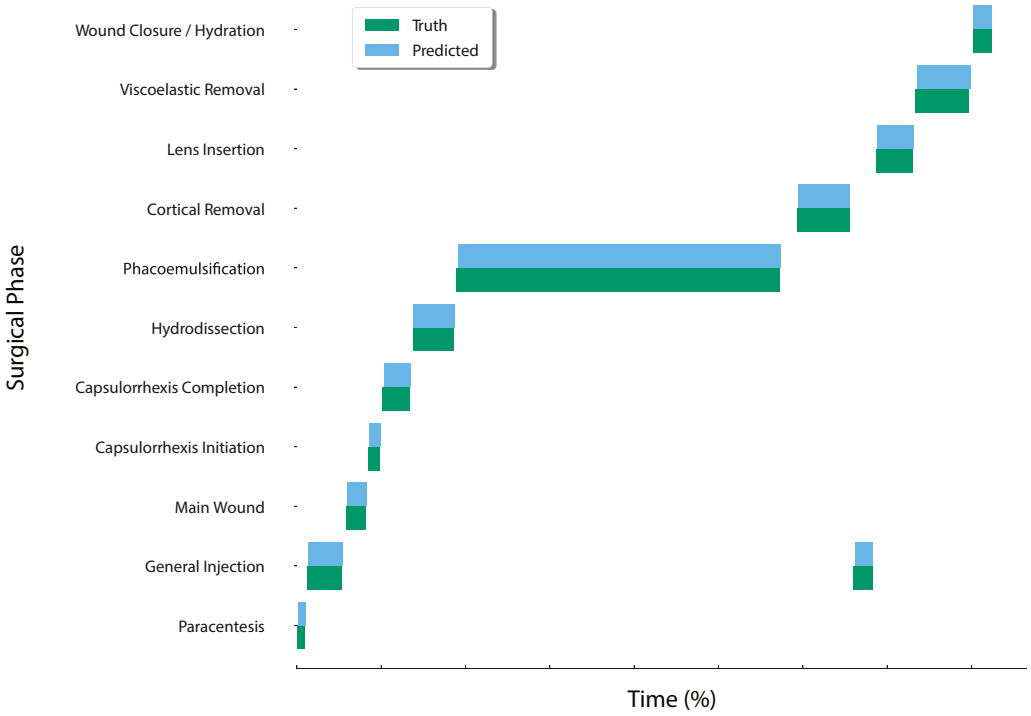

Figure 3. Sample Gantt chart for the final ensemble (CatStep) model. The chart depicts ground truth (green) and predicted (blue) phase identities as time progresses through the surgical video.

reported here demonstrates state-of-the-art performance for cataract surgical phase segmentation.

The BigCat dataset, employed in the training and evaluation of our models, offers an advantage over previously utilized datasets because of its significant size (nearly 4 million frames of surgical video). The variety, deep annotation, and careful curation of the BigCat dataset have likely contributed to the improvements in performance in surgical phase segmentation seen in this study.

In our study, the Densenet169 model demonstrated superior performance when compared to the I3D and CNN+RNN models on the majority of surgical phases. However, exceptions were observed in capsulorrhexis initiation and cortical removal, where the I3D model outperformed the Densenet169 by 7% and 2%, respectively. This discrepancy in performance may be attributed to the I3D model's inclusion of temporal features, which may aid in differentiating between the movements of surgical instruments that are either present in other phases (irrigation-aspiration handpiece) or spatially similar to other instruments (cystotome). It appears that the predictions for other phases are less reliant on temporal features and can be accurately predicted using only spatial features, as indicated by the Densenet169 model's strong performance. However, the Densenet was exposed to more individual instances of the data compared to the I3D model as the Densenet received individual frames rather than a sequence of 30 frames with a stride of 10 frames. Given that each model demonstrated strong performance in predicting different surgical phases, it was hypothesized that an Ensemble model incorporating features from both models would yield optimal results. This hypothesis was verified through the development of an Ensemble model and its evaluation on the validation set, where the Ensemble outperformed each individual model. Although the Ensemble model exhibited the highest performance, its processing time for videos was approximately 3 times longer than that of the Densenet at 0.054 seconds per frame. Applications focusing on real-time processing or deployment in resource-limited settings could reasonably consider the lighter-weight models presented here. Because the goal in this work was to optimize segmentation performance so as to reliably enable downstream analyses, however, the Ensemble model was selected as our final model (and named CatStep). The CatStep model achieved an F1-score of 0.91 and an AUROC of 0.95 on the hold-out testing set. These results demonstrate a significant improvement in surgical phase segmentation compared to previous studies. The AUROC of 0.95 of our CatStep model surpasses that of the model proposed by Yu et al,[2] which achieved an AUROC of 0.78 when trained on 100 cataract videos. The 0.99 accuracy in testing of CatStep also outpaced the more complex residual neural network + RNN combination proposed by Zisimopoulos et al,[3] which achieved an accuracy of 0.78 in their study.

To examine the influence of spatial features on performance, the Densenet was trained on multiple image resolutions and performance of the models was assessed. The Densenet trained on the highest resolution images (480 × 270) achieved a considerably higher F1-score (0.91) than those utilizing downsampled images (0.90 for 240 × 135 and 0.85 for 120 × 68). This appears to indicate the importance of spatial features in the identification and

segmentation of phases. Although the 240 × 135 resolution was ultimately chosen in this study to address space constraints, the results at higher resolution indicate a clear path toward higher performance in settings in which computational resources are less of a consideration.

In this work, we sought to focus not just on frame-level phase classifications but also phase-level segmentations. Automating the analysis of surgical phases requires the ability to predict start and end timestamps for each phase, rather than simply predicting the identity of a single frame at a time. Accordingly, we reported here several metrics on our model's phase-level prediction performance. CatStep achieved MAEs in phase start and end times of 2.3 seconds and 1.6 seconds, respectively. Median absolute error for phase start and end times were just 0.3 seconds and 0.1 seconds, respectively, indicating that outliers affected the MAEs. The ability to predict phase start and end times within 1 second of GT for the majority of cases indicates the applicability of CatStep to surgical analysis automation.

The final CatStep model had a segmental edit score of 0.92, indicating high performance in correctly identifying the relative ordering of phases. Furthermore, 31 out of 38 predicted surgical sequences in the testing set had ≤ 1 insertions or deletions, showcasing the high accuracy of the model in predicting surgical sequences. The final model's Over-Segmentation Score of 0.89 indicates infrequent fragmentation of phases. The segmental F1-score or F1@k was also used to assess the final model's performance on accurately segmenting phase boundaries. The model achieved F1@k scores of 0.95, 0.95, and 0.94 for k thresholds of 50, 60 and 70, respectively. The minimal reduction in F1-score as the IoU threshold increased indicates that there was a high degree of overlap between the predicted segmentation of the phases and the true segmentation.

Limitations of this study include using surgical videos from a single institution, although 9 different attending surgeons with varying techniques, illumination preferences, and instrumentation preferences were included in the dataset. Because of the sheer quantity of labeled frames (nearly 4 million) as well as the general difficulty in specifying the exact frame at which a phase transition occurs, a level of uncertainty regarding human-generated phase labels is expected at phase boundaries. Furthermore, inclusion of atypical cataract surgeries will help improve generalizability to complex cases and varying surgeon training levels. Future directions of this work will also include postsegmentation processing and analysis of each surgical phase identified so as to further expand systems for automated surgical analysis.

Footnotes and Disclosures

Originally received: July 15, 2023.
Final revision: September 6, 2023.
Accepted: September 15, 2023.
Available online: October 1, 2023. Manuscript no. XOPS-D-23-00174.

[1] Department of Ophthalmology and Visual Sciences, Kellogg Eye Center, University of Michigan, Ann Arbor, Michigan.

[2] School of Medicine, Wayne State University, Detroit, Michigan.

[3] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan.

[4] Department of Computer Science, University of Michigan, Ann Arbor, Michigan.

[5] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): B.T.: Grant—University of Michigan GME Innovations Fund, The Doctors Company Foundation.

N.N.: Grant—University of Michigan GME Innovations Fund, The Doctors Company Foundation, NIH K12EY022299.

HUMAN SUBJECTS:

Human subjects data were included in this study. This study received full ethics approval by the Michigan medicine institutional review board (HUM00160950), and it was determined that informed consent was not required because of its retrospective nature and the anonymized data utilized in this study. The study was carried out in accordance with the tenets of the Declaration of Helsinki. Animal subjects were not used in this study.

Author Contributions:

Conception and design: Mahmoud, Zhang, Matton, Mian, Tannen, Nallasamy

Analysis and interpretation: Mahmoud, Zhang, Matton, Tannen, Nallasamy

Data collection: Mahmoud, Zhang, Matton, Mian, Tannen, Nallasamy

Obtained funding: Tannen, Nallasamy

Overall responsibility: Mahmoud, Zhang, Matton, Mian, Tannen, Nallasamy

Abbreviations and Acronyms:
3D-CNN = 3-dimensional convolutional neural network; **AUROC** = area under the receiver operating characteristic curve; **CNN** = convolutional neural network; **GT** = ground truth; **I3D** = inflated 3-dimensional; **MAE** = mean absolute error; **ML** = machine learning; **NN** = neural network; **RNN** = recurrent neural network; **SES** = segmental edit score.

Keywords:
AI, Cataract surgery, Machine learning, Resident training, Surgical feedback.

Correspondence:
Nambi Nallasamy, MD, Kellogg Eye Center, University of Michigan, 1000 Wall St, Ann Arbor, MI 48105. E-mail: nnallasa@med.umich.edu.

References

1. Randleman JB, Wolfe JD, Woodward M, et al. The resident surgeon phacoemulsification learning curve. *Arch Ophthalmol*. 2007;125:1215–1219.
2. Yu F, Croso GS, 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:e191860.
3. Zisimopoulos O, Flouty E, Luengo I, et al. DeepPhase: surgical phase recognition in CATARACTS videos. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes*

in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol 11073 LNCS. New York: Springer Verlag; 2018:265–272.
4. Garcia Nespolo R, Yi D, Cole E, et al. Evaluation of artificial intelligence–based intraoperative guidance tools for phacoemulsification cataract surgery. *JAMA Ophthalmol*. 2022;140:170–177.
5. Sahu M, Szengel A, Mukhopadhyay A, Zachow S. Surgical phase recognition by learning phase transitions. *Curr Dir Biomed Eng*. 2020;6:20200037.
6. Zhang Y, Bano S, Page AS, et al. Large-scale surgical workflow segmentation for laparoscopic sacrocolpopexy. *Int J Comput Assist Radiol Surg*. 2022;17:467–477.
7. Matton N, Qalieh A, Zhang Y, et al. Analysis of cataract surgery instrument identification performance of convolutional and recurrent neural network ensembles leveraging BigCat. *Transl Vis Sci Technol*. 2022;11:1–8.
8. Quellec G, Lamard M, Cochener B, Cazuguel G. Real-time segmentation and recognition of surgical tasks in cataract surgery videos. *IEEE Trans Med Imaging*. 2014;33:2352–2360.
9. Al Hajj H, Lamard M, Charriere K, et al. Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network. *Proc Annu Int Conf IEEE Eng Med Biol Soc*. 2017;2017:2002–2005.
10. Schoeffmann K, Taschwer M, Sarny Klinikum Klagenfurt S, et al. Cataract-101-video dataset of 101 cataract surgeries. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. New York, NY: Association for Computing Machinery; 2018:421–425.
11. 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:e191860.
12. Zang D, Bian G-B, Wang Y, Li Z. An Extremely Fast and Precise Convolutional Neural Network for Recognition and Localization. In: *of Cataract Surgical Tools, Lecture Notes in Computer science*. Cham, Switzerland: Springer; 2019.
13. Morita S, Tabuchi H, Masumoto H, et al. Real-time surgical problem detection and instrument tracking in cataract surgery. *J Clin Med*. 2020;9:3896.
14. Al Hajj H, Lamard M, Conze P-H, et al. CATARACTS: challenge on automatic tool annotation for cataRACT surgery. *Med Image Anal*. 2019;52:24–41.
15. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*. New York, NY: IEEE; 2017:4724–4733. https://doi.org/10.1109/CVPR.2017.502.
16. Gammulle H, Ahmedt-Aristizabal D, Denman S, Tychsen-Smith L, Petersson L. *Fookes C. Continuous Human Action Recognition for Human-Machine Interaction: A Review*. New York, NY: Association for Computing Machinery; 2023.
17. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD. *Temporal Convolutional Networks for Action Segmentation and Detection*. New York, NY: IEEE; 2017.
18. Lea C, Vidal R, Hager GD. Learning convolutional action primitives for fine-grained action recognition. https://github.com/colincsl/LCTM. Accessed February 9, 2023.