Performance Evaluation in Cataract Surgery with an Ensemble of 2D-3D Convolutional Neural Networks[1*]

Ummey Tanin[1], Adrienne Duimering[2], Christine Law[2], Jessica Ruzicki[2], Gabriela Luna[2] and Matthew Holden[1]

[1] School of Computer Science, Carleton University, Ottawa, ON, Canada
ummeytanin@cmail.carleton.ca, matthew.holden@carleton.ca

[2] Dept. of Ophthalmology, School of Medicine, Queen's University, Kingston, ON, Canada
aduimering@qmed.ca, christine.law@queensu.ca,
j.ruzicki@me.com,gaby.lahaie@gmail.com

Abstract. An important part of surgical training in ophthalmology is understanding how to proficiently perform cataract surgery. Operating skill in cataract surgery is typically assessed by real-time or video-based expert review using a rating scale. This is time-consuming, subjective and labour-intensive. A typical trainee graduates with over 100 complete surgeries, each of which requires review by the surgical educators. Due to the consistently repetitive nature of this task, it lends itself well to machine learning based evaluation. Recent studies utilize deep learning models trained on tool motion trajectories obtained using additional equipment or robotic systems. However, the process of tool recognition by extracting frames from the videos to perform phase recognition followed by skill assessment is exhaustive. This project proposes a deep learning model for skill evaluation using raw surgery videos that is cost-effective and end-to-end trainable. An advanced ensemble of convolutional neural network models is leveraged to model technical skills in cataract surgeries and is evaluated using a large dataset comprising almost 200 surgical trials. The highest accuracy of 0.8494 is observed on the phacoemulsification step data. Our model yielded an average accuracy of 0.8200 and an average AUC score of 0.8800 for all four phase datasets of cataract surgery proving its robustness against different data. The proposed ensemble model with 2D and 3D convolutional neural networks demonstrated a promising result without using tool motion trajectories to evaluate surgery expertise.
Keywords: Surgical Skill Assessment ·Deep Learning ·2D-3D CNN.

1 Introduction

Cataract surgery is the most common procedure in ophthalmology which involves the removal of the crystalline lens of the eye that opacifies over time which leads to

[1*] This work was partially supported by the Natural Science and Engineering Research Council (Grant RGPIN-2020-05582), Compute Ontario and the Digital Research Alliance of Canada.

reversible visual impairment and is replaced with an artificial lens [1]. The crystalline lens of the human eye refracts light to focus a clear image on the retina.

All ophthalmology residents are trained by faculty surgeons to proficiently perform the procedure of cataract surgery and learn the various surgical steps, tools and techniques. Although there may be slight variations in the surgical tools and techniques, the actual steps of the surgery are fairly universal. A trainee surgeon may display the medical knowledge of how to perform the surgical steps; however, intraoperative skills using microsurgical instruments, fine dexterity, and microscope control requires additional assessment that cannot be based on knowledge alone.

Skill assessment can be performed by analyzing the intraoperative surgical videos conducted by the trainee surgeons. Many rating scales established for evaluating surgical skills depend on the subjective interpretation and opinion of the faculty surgeons appointed to observe the surgery. For example, ICO-OSCAR [12] is a validated rating scale where an assessor examines the intraoperative surgery based on predefined steps, categories and scale descriptions. The human graders mark the performance of the trainee following the assessment criteria; however, some of the descriptors can be interpreted in different ways which involve a subjective grading component. The implication is that cataract surgical training can be time-consuming, labor-intensive, and prone to human error. Thus, there is a need for the objective evaluation of technical skills apart from subjective measures.

Video classification algorithms can be applied to efficiently analyze surgical videos recorded for educational purposes instead of manual analysis. Convolutional neural network (CNN) models have been proven to be effective in identifying actions in videos [2][3][4]: this inspires this work to use action recognition models for skill assessment. The surgical instruments in microscope videos are tiny compared to objects in public video datasets such as UCF-101, where objects appear in an identifiable and larger shape. Thus, state-of-the-art video classification models might not perform well in evaluating microscope videos. Scholars have tried to perform classification on publicly available cataract datasets that were published in the Cataract Grand Challenge competitions [5]. Videos in these datasets are well recorded with good lighting conditions, lens focus and HD resolution, which is not necessarily possible in locally collected real-time surgeries. Authors in [6] have reported that the published surgical tool detection models do not generalize well on the datasets collected from local hospitals. Thus, the publicly available pre-trained neural network model may not work well on raw cataract surgery videos.

Most prior work on surgical video analysis is dedicated to surgical workflow analysis and skill assessment by performing automated recognition of instruments and surgical phases, tool usage and tool movement with different pre-trained 2D CNN's [6][7][8][14][18][19]. Some earlier work uses 3D CNNs for skill categorization [15][9]. Frequency analysis or motion analysis of surgical activities has also been used for encoding motion features from surgery videos [16][17]. Although some work (e.g. [14],

[19]) investigated real surgery video clips using 2D CNN, the majority of them rely on simulated surgery video collected from robotic systems. In summary, few prior algorithms were designed using real surgery data and none of the works deployed an ensemble of 2D and 3D CNNs for predicting classification scores on raw cataract surgery data. Kim et al. performed an objective assessment of technical skills on the videos of cataract surgeries using information about instrument usage (position or velocity) such as tool trajectories as a representation of a video frame [9]. Alternatively, they propose an approach using optical flow encodings to represent video frames. The authors reported a higher accuracy of 84% using tool tip velocity information; their approach using an optical flow representation computed directly from video data achieved an accuracy of only 63%. To obtain tool motion information, trajectories of surgical instruments were determined from crowd-sourced annotations. This information, however, requires additional equipment for tool tracking or manual annotation. This is not feasible for most skill assessment setups.

This work presents a convolutional neural network model to categorize surgeons as novices or experts and thus systematize the process of grading surgical competence in raw cataract surgery videos. The proposed model is cost-effective and end-to-end trainable. This tool is intended to assist human raters in grading by producing highly accurate results.

2 Methods

2.1 Dataset Preparation and Preprocessing

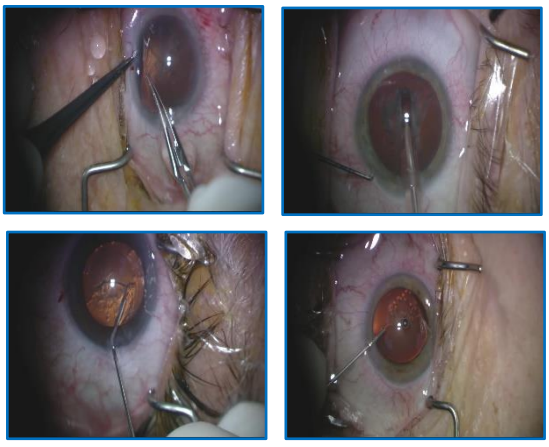

Fig. 1. Phases in the cataract surgery dataset. Video frames from Capsulorrhexis formation (upper left), Phacoemulsification (upper right), Hydrodissection (lower left), and Viscoelastic insertion (lower right).

The dataset consists of 197 recordings of microscopic video of varying lengths with a frame rate of 29 fps and a resolution of 1920 × 1080 recorded during cataract surgery [8]. These videos were collected as part of the routine clinical workflow. Only recordings that involved multiple surgeons, were, incomplete or failed were excluded. There were seven faculty surgeons and five trainee surgeons who participated in the study for recording surgical procedures. The dataset has skill-level annotations, indicating whether each surgery was performed by a novice or an expert. This was done by appointment status. Fourth-year and fifth-year residents, who require direct supervision when performing surgery, were considered novices. Attending surgeons, who do not require supervision during surgery, were considered experts. While there are 13 phases annotated in the dataset, this work utilizes data from four phases which can be considered the most salient steps for identifying operative competence in cataract surgery: Capsulorrhexis formation, Phacoemulsification, Hydrodissection and Viscoelastic insertion (Figure 1). Several steps are followed to prepare the dataset, as described below.

1. Videos of each surgical phase are trimmed from the raw surgery video using the phase's start time and end time indicated in the annotations.
2. Each video was segmented into video clips with a duration of 16 seconds. Each of these video clips is treated as a single data sample representing the original phase-wise video. The label of each video clip is based on the appointment status of the surgeon who performed the surgery.
3. Each video is downsampled to 1 fps, such that there are 16 frames in each 16s video.
4. Temporal padding of the nth frame in a video clip is performed so that each video sample can be represented using 16 frames only. This is necessary due to the varying lengths of each surgical phase across patients.
5. The 16 frames are extracted from each video snippet and stored as images in an individual folder for each video.

Original videos of the expert class were shorter in length leaving fewer snippets in the data category than the novice recordings, so the samples were up-sampled by applying a random rotation of up to 5 degrees to handle data imbalance.

2.2 Ensemble Model for Skill Assessment

Model Ensembling is often considered a powerful technique to form a noise-invariant and robust model. We deployed an ensemble of two deep 2D CNN-LSTM models and one 3D CNN model for performing skill assessment in surgical videos (Figure 2). We utilized model ensembling with a basic bagging approach where separate models work on the same training set parallelly and the outputs from the last layer of each model are concatenated to produce the final prediction scores. In the ensemble model, all the models learn different features from the same input, which helps to reduce the variance and improve the performance of video classification to a greater extent. The proposed ensembling approach is highly inspired by C3D [20] and more precisely T3D [21] who

have fused features from both 2D and 3D convolutional networks to design a competent video classification model that has outpaced many state-of-the-art models. The first model is a two-dimensional CNN model where convolution is applied in a time-distributed manner. The second model is developed employing the first model as a baseline. These 2D CNN models can learn spatial features. We use a shared LSTM layer between these two models to analyze the frame-level features temporally. Finally, we integrate a simple 3D CNN model with the first two models, designed using 3D convolutional layers.

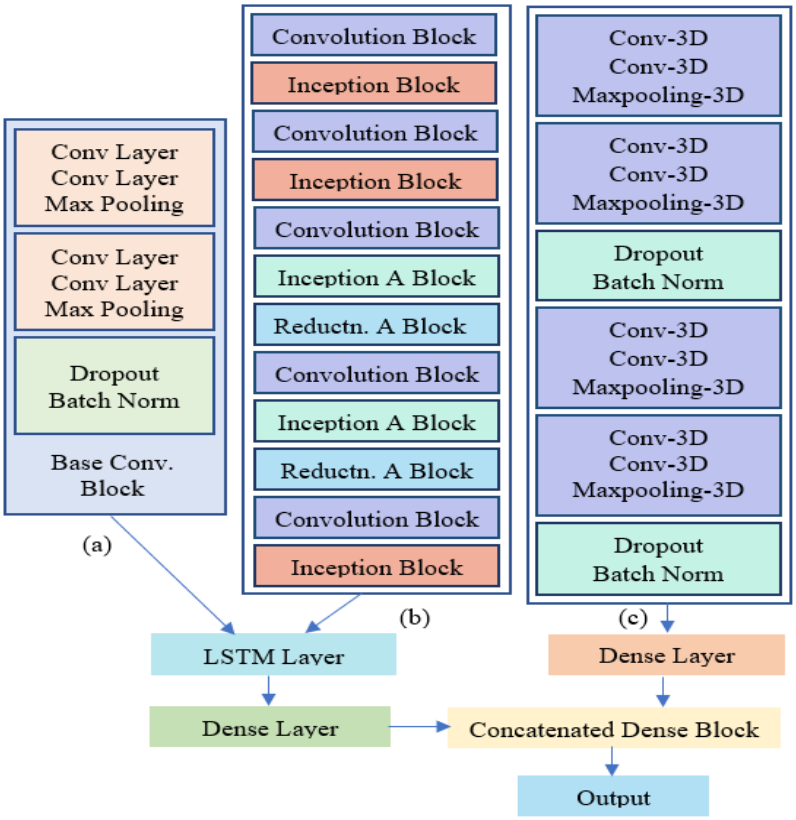

Fig. 2. Ensemble Model ; (a): 2D CNN LSTM, (b): 2D CNN LSTM with Inception V4 Modules, (c): 3D CNN.

A sequence of 16 frames from each input video is fed to both the 2D CNN-LSTM models and the 3D CNN as a separate input. An earlier work uses two separately trained networks for modeling spatial and temporal features individually [2]; however, this approach needs the optical flow images to be pre-computed and it requires the RGB videos to be incorporated with the optical flow images as the input to the models. This

strategy has excessive memory needs which is impractical when a large dataset is utilized. Therefore, our work does not utilize separately computed optical flow images so that the network can be made end-to-end trainable and efficient.

2.3 Basic 2D CNN with LSTM

This model is designed using time-distributed convolutional layers along with a filter size of 3×3 that allows a sequence of frames to be processed by a 2D CNN**.** The same base convolutional block is repeated a few times to deepen the CNN structure (see Fig. 2(a)). The Dropout layer is associated with each batch norm layer as an advanced regularization approach. Features learned from the last convolutional block of this model are passed down to the concatenation layer before it is fed to a shared LSTM layer for analyzing the temporal aspects.

2.4 2D CNN with Inception V4 and LSTM

This model is built with some modifications to the first 2D CNN model (see Fig. 2(b)). An Inception block is incorporated between each time-distributed convolutional layer for performing feature extraction which is inspired by Inception V4 [10]. We have utilized the basic inception module, Inception A module and Reduction module from Inception V4 which are designed to resolve the issue of computational expenses while efficiently learning patterns at various scales using different filters. The output layer after the last inception block in Figure 2(b) is concatenated with the output layer from the first model. This concatenated output is then fed to a shared LSTM layer.

2.5 3D CNN

The 3D CNN contains similar convolutional blocks as the first 2D CNN-LSTM where 3D layers are used instead of 2D layers (see Fig. 2(c)). They are like the 2D CNNs, where features from both spatial and temporal dimensions are extracted by performing 3D convolutions with filter size of 3×3.

3 Implementation Details

The proposed model is implemented in TensorFlow on Digital Research Alliance of Canada Graham cluster. The author's thesis explains the finest details about the implementation [13]. Greyscale images were used with a resolution of 128 × 128. ReLU is used as an activation function while sigmoid activation is used in the output layer only. The single LSTM has 64 nodes with a dropout value of 0.2. RMSprop is used as an optimizer function with a learning rate of 0.0001 and binary cross-entropy being the loss function. The ensemble model was trained end-to-end, and the following parameter space was searched: using 100 - 200 epochs, batch sizes of 4, 6 and 8, and a dropout probability ranging from 0.2 - 0.3 for the convolution layers.

We experimented on three model configurations. First, we use an L2 kernel regularizer with a value of 0.001 to the kernel to avoid overfitting. Second, we add batch normalization with dropout following all convolution blocks to improve the learning mechanism of the model following the idea of Chen et al. [11]. We opted to augment the data such that the model could better learn patterns in the data without exaggerating the memory requirements. To this end, we compute and apply HOG descriptors for image frames to the image sequences. This augments spatial features in the frames with minimal extra computation resources. HOG is a feature descriptor that emphasizes the shape of any object in a given image and hence augments the spatial features in video frames which is crucial for microscopic recordings.

A five-fold trial-out cross-validation strategy was used with the same data partitions for all splits (i.e. at each iteration 80% of the data is used for training and validation, while 20% of the data is used for testing). Subsequently, we have also performed a leave-one-surgeon-out cross-validation strategy and have reported the result. This evaluates how well the model generalizes to never-before-seen surgeons. The reported accuracies and other performance metrics were averaged over all folds.

4 Results and Discussion

The model for surgical skill assessment is trained and evaluated using video data from four cataract phases. The model classifies the video data as either expert or novice. First, the validation accuracy of the model is computed using Capsulorrhexis data for all the model configurations and hyperparameters discussed above. Table 1 illustrates the performance of each component of the ensemble. Table 2 shows performance under various model configurations. As seen in Table 2, the baseline training accuracy is 95.0% which is higher than the 77.8% validation accuracy. L2 regularization reduces the generalization gap between the accuracy scores. It makes the model more stable and reduces the memory requirements during training where the total number of model parameters changes from 18.6 million to 12 million. Interestingly, combining batch normalization and dropout in the convolution blocks also improved the model's performance, and the model performance was further improved by applying HOG filters on the image frames.

Table 1. Scores for all three models

Model Names	2D CNN	2D CNN- Inception	3D CNN
Train Acc.	0.9400	0.9571	0.9608
Valid Acc.	0.7500	0.7819	0.7405

Table 2. Model scores with different settings

Approach	Baseline	L2 Regularizer	Batch Norm + Dropout	HOG Filter
Train Acc.	0.9500	0.9681	0.9685	0.9999
Valid Acc.	0.7777	0.8214	0.8400	0.8900

Following these experiments, once the model configuration and hyperparameter settings are finalized, the test accuracy is computed for all four phases using the optimal configuration of the model. The same model gets trained individually using data from each phase following the same optimal configuration. The average classification results for all phases are demonstrated in Table 4 as an average over all cross-validation folds. The data in all four surgical phases are different as they have different video duration and include dissimilar surgical steps and tools. The videos for Capsulorrhexis are on average 156.3 seconds while for Phacoemulsification they are 729.5 seconds on average. The validation and test accuracy for the other 3 phases were consistent with the model accuracy for Capsulorrhexis data (Table 3). The model showed notable results with an average AUC score of 88.00% and sensitivity score of 85.80% over all data for all phases (Table 3,4). The consistency in the overall results for all phases implies that the model is robust to different types of surgical phases.

Table 3. Skill assessment scores on all four phases

Phase Name	Train Acc	Val. Acc.	Test Acc.	Sensitivity	AUC
Capsulor.	0.9999	0.8960	0.8100	0.8340	0.8700
Phacomul.	1.0000	0.9400	0.8494	0.8833	0.9000
Hydrodiss.	0.9999	0.9072	0.8130	0.8473	0.8820
Viscoelast.	1.0000	0.9100	0.8300	0.8677	0.8897

Table 4. Test accuracy for skill assessment on all five folds for all phases

Phase Name	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg.
Capsulor.	0.8700	0.8200	0.8000	0.7900	0.7700	0.8100
Phacomul.	0.8100	0.8072	0.8000	0.9600	0.8700	0.8494
Hydrodiss.	0.7760	0.7590	0.9000	0.8000	0.8300	0.8130
Viscoelast.	0.8200	0.8202	0.8000	0.8300	0.8800	0.8300

Table 5. Skill assessment scores on surgeon-out cross-validation

Surgeon Id	A	B	C	D	E	F	G
Test Acc	0.7502	0.7203	0.7789	0.6917	0.6910	0.7740	0.7423

The result from the surgeon out cross-validation on capsulorhexis data is depicted in Table 5. However, the results from the model are not as good as they were for trial-out cross-validation in Tables 2, 3 and 4 since the number of surgeons who participated in the surgery recordings was only 12.

For further analysis of how the model results can be exploited to assist surgical evaluation, the real-time prediction result from a surgical video is depicted in Fig. 3.

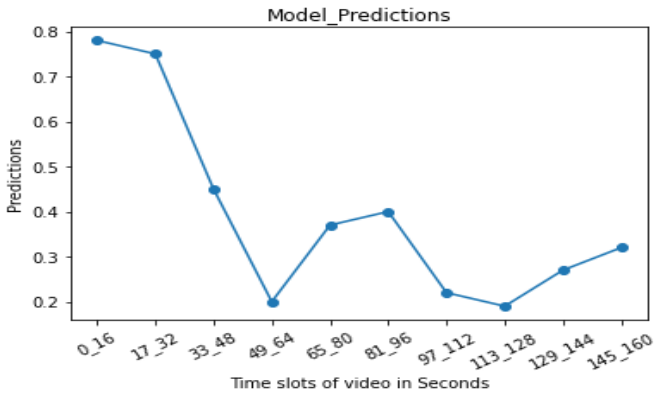

Fig. 3. Real-time prediction on an unseen video sample identifying Novice

A 147-second video was trimmed into 16-second clips and video duration was plotted in the x-axis as 16-second snippets. In Figure 3, it is visible that for the first two video clips (video clip duration 0-16s and 17-32s) the prediction score is higher than 0.5 identifying an expert-level (Expert Score is 1) surgical performance; however, for subsequent video clips, the score is within 0.30-0.45 which implies the surgical performance during those times was at the novice level (Novice Score is 0). This prediction can be utilized to identify exactly at what time the surgeon showed novice-level proficiency during the surgery, and thus, what parts of the surgery the trainee must practice more to achieve proficiency.

The outcome of our work indicates a substantial improvement over earlier results for skill assessment. Earlier research work reported a similar accuracy for skill classification by utilizing motion trajectories obtained from tool tip positions in the surgery videos [9]. However, their video-level accuracy was 63% only. Therefore, an average testing accuracy of 82.6% is adequate given the fact that the proposed model can be even better than the human rating consistency (which was found to be 71% on a separate dataset) [13]. Our model yields reasonable accuracy for skill assessment on four different phases on a larger dataset, and it does not need tool trajectories to be computed from raw surgeries for predicting operator proficiency. Another skill analysis was conducted on our dataset in an earlier experiment using a late supervision approach [8]. The highest skill classification accuracy achieved from the earlier work was 63.3%.

One of the primary limitations of this project is the ground-truth annotation for skill level. These annotations do not provide fine-grained information about skill level. It is also possible that some of the staff or faculty surgeons might not truly be experts, and some residents may be very proficient toward the end of their training. Another limitation is the small size of the dataset (only seven experts and five novices participated). Finally, the model is not designed to predict skill classification scores on an entire surgery video.

5 Conclusion

The performance of the ensembled model reveals that the model is effective in capturing intraoperative skills in cataract surgeries without human grading input. An ensemble of 2D and 3D convolutional networks with different sizes of convolution filters can successfully study technical skills in cataract surgeries without using tool motion trajectories for evaluating expertise in surgeries. Histogram descriptors turned out to be effective for augmenting visual appearance in the data which improved the model's ability to assess surgical proficiency with an increase in accuracy. One of the noteworthy achievements of this work is obtaining consistent results from four phases of cataract surgery. This work shows the feasibility of deployment as a tool in the certification of cataract surgery proficiency, where a trainee has their learning curves monitored using the proposed model prior to final proficiency evaluation by an expert preceptor.

Future work involves predicting a more granular skill score such as skill level on a rating scale (e.g. ICO-OSCAR), collecting data from more unique surgeons and from more than one hospital to formulate a more generalized algorithm for surgical skill recognition, and developing a video classifier that can predict skill level from an entire surgery video.

6 References

1. Davis G.: The evolution of cataract surgery. Missouri Medicine, vol. 113(1), pp. 58–62 (2016).
2. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S. : Behavior recognition via sparse spatial-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72 (2005).
3. Jeff, D., Lisa, H., Subhashini, G., Marcus, R., Sergio, G., Kate, S., Trevor, D.: Long-term recurrent convolutional networks for visual recognition and description. CVPR 2015, pp. 2625–2634.
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, http://www.deeplearningbook.org (2016).
5. Padoy, N. , Twinanda, A. P.: M2CAI 2016 Challenge, http://camma.u-strasbg.fr/ m2cai2016/.
6. Sokolova, N., Schoeffmann, K., Taschwer, M., Putzgruber-Adamitsch, D., ElShabrawi, Y.: Evaluating the generalization performance of instrument classification in cataract surgery videos. In: Conference on Multimedia Modeling, pp. 626–636,(2019).
7. Sahu, A. S. M., Mukhopadhyay, A., Zachow, S.: Tool and phase recognition using contextual cnn features. M2CAI 2016, (2016).
8. Ruzicki, J., Holden, M., Cheon, S., Ungi, T., Egan, R. and Law, C.: Use of machine learning to assess cataract surgery skill level with tool detection. Ophthalmology Science, vol.3, pp.100235,(2022)

9. Kim, T. S., O'Brien, M., Zafar, S., Hager, G., Sikder, S., Vedula, S. S.: Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. International Journal of Computer Assisted Radiology and Surgery, vol.14, pp.1097-1105, (2019).
10. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A. A.: Inception-v4, inception resnet and the impact of residual connections on learning. In: Association for the Advancement of Artificial Intelligence (AAAI), (2017).
11. Chen, G., Chen, P., Shi, Y., Hsieh, C.-Y., Liao, B., Zhang, S.: Rethinking the usage of batch normalization and dropout in the training of deep neural networks. ArXiv, vol. abs/1905.05928, (2019).
12. Surgical Assessment Tool: ICO-OSCAR, https://icoph.org/?s=ICO-OSCAR.
13. Tanin, U.H. : Deep Video Analysis Methods for Surgical Skill Assessment in Cataract Surgery. Master's Thesis, Carleton University (2022). https://curve.carleton.ca/77d85ffd0cfb-468b-b9f0-81939778349f
14. Padoy, N., Twinanda, A. P. M2CAI 2016 Challenge, http://camma.u-strasbg.fr/ m2cai2016.
15. Funke, I., Mees, S., Weitz, J., Speidel, S.: Video-based surgical skill assessment using 3D convolutional neural networks. International Journal of Computer Assisted Radiology and Surgery, vol. 14, doi. 10.1007/s11548-019-01995-1, (2019).
16. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E. L., Clements, M. A., Essa, I.: Automated assessment of surgical skills using frequency analysis. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, vol. 9349, pp. 430-438. MICCAI 2015, doi. https://doi.org/10.1007/978-3-319-24553-9_53, Berlin, Heidelberg (2015).
17. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E. L., Essa, I.: Video and accelerometer-based motion analysis for automated surgical skills assessment. In: International Journal of Computer Assisted Radiology and Surgery, vol. 13, pp. 443-455 (2018).
18. Zia, A., Guo, L., Zhou, L., Essa, I., Jarc, A.M.: Novel evaluation of surgical activity recognition models using task-based efficiency metrics. In: International Journal of Computer Assisted Radiology and Surgery, vol. 14, pp. 2155 – 2163 (2019).
19. Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., and Fei-Fei, L.: Tool Detection and Operative Skill Assessment in Surgical Videos Using Region-Based Convolutional Neural Networks. IEEE Winter Conference on Applications of Computer Vision, (2018).
20. Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497 (2015).
21. Diba, A., Fayyaz, M., Sharma, A., Karami, A. H., Arzani, M. M., Yousefzadeh, R. and Gool, L. V.: Temporal 3d convnets: New architecture and transfer learning for video classification. ArXiv, vol. abs/1711.08200 (2017).