International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-025-03366-5

ORIGINAL ARTICLE

Objective skill assessment for cataract surgery from surgical microscope video

Rebecca Hisey[1] · Henry Lee[1] · Adrienne Duimering[2] · John Liu[2] · Vasudha Gupta[2] · Tamas Ungi[1] · Christine Law[2] · Gabor Fichtinger[1] · Matthew Holden[3]

Received: 28 November 2024 / Accepted: 21 March 2025

Abstract

Objective Video offers an accessible method for automated surgical skill evaluation; however, many platforms still rely on traditional six-degree-of-freedom (6-DOF) tracking systems, which can be costly, cumbersome, and challenging to apply clinically. This study aims to demonstrate that trainee skill in cataract surgery can be assessed effectively using only object detection from monocular surgical microscope video.

Methods One ophthalmologist and four residents performed cataract surgery on a simulated eye five times each, generating 25 recordings. Recordings included both the surgical microscope video and 6-DOF instrument tracking data. Videos were graded by two expert ophthalmologists using the ICO-OSCAR:SICS rubric. We computed motion-based metrics using both object detection from video and 6-DOF tracking. We first examined correlations between each metric and expert scores for each rubric criteria. Then, using these findings, we trained an ordinal regression model to predict scores from each tracking modality and compared correlation strengths with expert scores.

Results Metrics from object detection generally showed stronger correlations with expert scores than 6-DOF tracking. For score prediction, 6-DOF tracking showed no significant advantage, while scores predicted from object detection achieved significantly stronger correlations with expert scores for four scoring criteria.

Conclusion Our results indicate that skill assessment from monocular surgical microscope video can match, and in some cases exceed, the correlation strengths of 6-DOF tracking assessments. This finding supports the feasibility of using object detection for skill assessment without additional hardware.

Keywords Video-based skill assessment · Object detection · Cataract surgery

Introduction

Traditionally, cataract surgery skill assessment relied heavily upon subjective expert supervision. Surgical residents were assigned mentors and assessed by experienced surgeons. This apprenticeship approach made it difficult to clearly evaluate skill development progress. Over the years, many skill assessment rubrics have been developed to increase objectivity of expert supervision [1–3]. These rubrics divide the procedure into multiple sections and provide descriptors to classify each performance level. Such rubrics serve to guide expert supervision and reduce subjectivity. Despite improvements, differences in experience between supervising experts may still generate interobserver variability [4, 5]. There is additional concern regarding the resource intensity of expert supervision [6, 7]. Due to the scarcity of instructors relative to the number of trainees, expert supervision reduces the opportunities for a trainee to obtain quality feedback in response to their surgical performance to only those times when an expert is available to observe. Therefore, it is preferable to automate skill assessment to improve scalability and reliability of cataract surgery education. Automating skill assessment also has the potential to eliminate explicit favoritism and minimize the effect of unconscious bias against individuals based on factors such as their gender, age, or skin color.

✉ Rebecca Hisey
rebecca.hisey@queensu.ca

1 School of Computing, Queen's University, Kingston, ON, Canada

2 Department of Ophthalmology, Queen's University, Kingston, ON, Canada

3 School of Computer Science, Carleton University, Ottawa, ON, Canada

Published online: 25 April 2025

There have been attempts to automate skill assessment through techniques such as using crowd sourcing [8]. While these approaches work reasonably well, they require the wide distribution of potentially sensitive data, which may lead to security concerns. Most automated methods for skill assessment evaluate surgical proficiency based on quantitative metrics of tool handling. Current gold-standard systems utilize conventional marker- or sensor-based tracking systems, such as optical, infrared, and electromagnetic (EM) tracking, to track tool movement. These systems can provide precise six-degree-of-freedom (6-DOF) pose information. Quantitative assessment metrics are calculated from this data by calculating metrics such as tool velocity, position, path length, and usage time. These metrics have been widely demonstrated to differentiate between trainee skill levels [9–11]. A recent study using EM tracking has also demonstrated that the rate of change in orientation may also be indicative of surgeon skill [12]. Several other approaches based on energy, force, and vibration sensing have also been tested [13–15], though this research mostly focuses on laparoscopic or robotic surgery and largely ignores open and microscopic procedures.

Despite the potential in these approaches, the use of physical sensors to assess trainee skill is a major barrier in many aspects. First, they significantly add to system costs and complexities, with most systems costing thousands of dollars or more depending on the modality. Even if we consider low-cost tracking systems such as ArUco [16], these systems still limit face validity by altering both the appearance and the physical interaction of the surgical tools, thereby impacting the trainee's overall experience. Finally, and perhaps most importantly, the use of 6-DOF tracking systems limits the predictive validity of automated skill assessment by precluding direct measurement of how competence translates from simulated learning to the clinical setting for procedures that are not minimally invasive or performed under robotic guidance.

Object detection, on the other hand, is capable of providing tracking-like information by identifying object locations in sequential frames. Object detection can be done using a consumer-grade camera and has the benefit of not requiring any modifications to the surgical tools. In other applications, object detection has successfully been shown to distinguish between levels of trainee skill using only a webcam [17]. Furthermore, in a recent study that examined the feasibility of assessing skill from two-dimensional data, it was demonstrated that two-dimensional data is comparable to three-dimensional data when used to perform skills assessment [18]. In the case of cataract surgery, there are no additional hardware requirements, as modern surgical microscopes typically include integrated digital video capabilities and can readily connect to computers for video recording and review. Importantly, this approach mirrors how human graders typically evaluate performance—by observing videos captured from the surgical microscope perspective—thereby aligning the automated analysis with the natural viewpoint of expert evaluators. In 2019, Kim et al. demonstrated that trainee skill could accurately be predicted from the surgical microscope video alone for the capsulorhexis steps of cataract surgery and found that using information about the specific tool movement yielded a better prediction of skill when compared to using the images or general image features such as optical flow patterns [19]. In addition, by analyzing specific tool motion, there is the added benefit of explainability. Deep neural networks are known to be somewhat of a black box, and it can be difficult to determine how the networks are making their predictions. By dividing skill assessment into two phases: detection of tool motion and skill prediction, we aim to provide clinicians with a better understanding of how scoring metrics were predicted.

In this study, we compare skill assessment predictions generated using object detection from surgical microscope video with those obtained from 6-DOF optical tracking of surgical tools, specifically using ArUco markers, which is currently the best low-cost option. Our goal is to determine whether object detection, using metrics derived from standard surgical videos, can provide a comparable or improved measure of trainee skill, offering a more accessible and less intrusive alternative to traditional tracking systems.

Methods

Experimental setup

The hardware setup for this experiment is shown in Fig. 1. The setup consisted of a simulated eye, six surgical tools, and a surgical microscope, along with a laptop computer and an Intel RealSense depth camera (www.intelrealsense.com). The surgical microscope recorded video at a resolution of 1260 x 720 pixels and a frame rate of six frames per second. The Intel RealSense camera was used exclusively to obtain 6-DOF tracking data by attaching ArUco markers to the six surgical tools: forceps, diamond keratome (straight), viscoelastic cannula, cystotome needle, diamond keratome (angled), and capsulorhexis forceps, as shown in Fig. 2. An additional reference marker was fixed on the workbench adjacent to the simulated eye. These markers were attached to the ends of the tools and were not visible in the surgical microscope video. Both the surgical microscope video and the 6-DOF tracking information were synchronized and streamed using the PLUS toolkit (https://plustoolkit.org). Data from the RealSense camera consisted only of positional tracking information; no video was recorded from this device.

The zoom on the microscope was kept constant across all trials, and all data was recorded using the open-source plat-

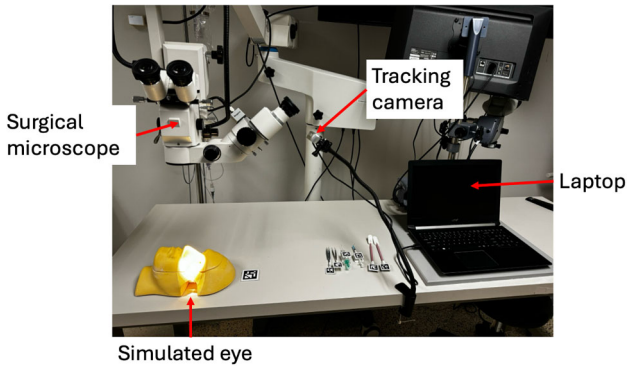

Fig. 1 Experimental setup for simulated cataract surgery

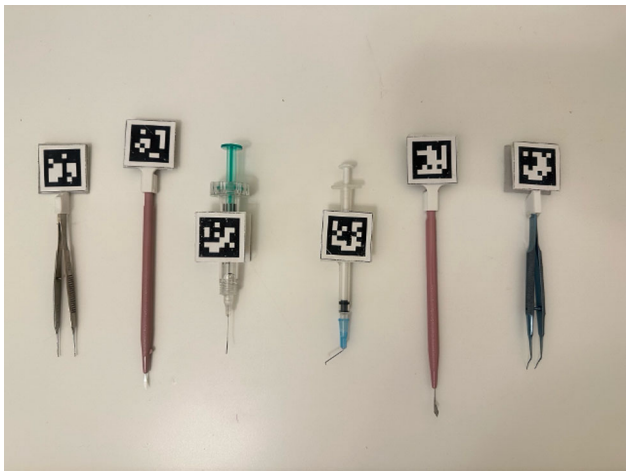

Fig. 2 Surgical tools used in cataract surgery, fitted with ArUco markers

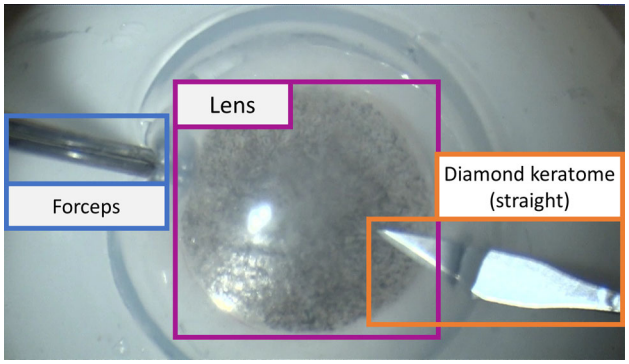

Fig. 3 Sample object detection annotations from surgical microscope video

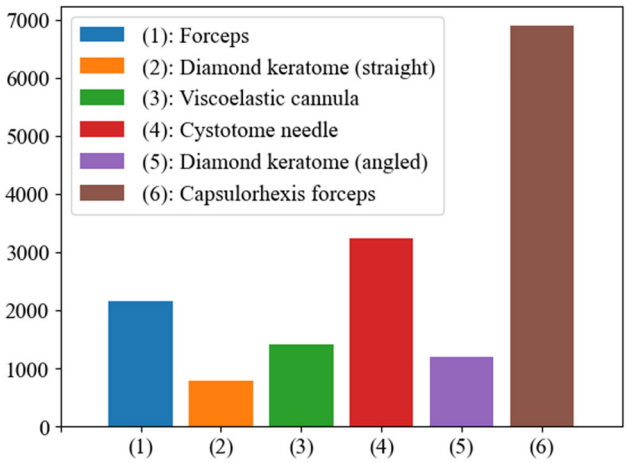

Fig. 4 Number of object detection annotations for each tool

form 3D Slicer (www.slicer.org). Although the frame rate of six frames per second is relatively low due to limitations of the ArUco tracking software and the laptop used for recording, graders reported no difficulty in evaluating the surgical videos. We focused on ArUco tracking as it is one of the most affordable options for 6-DOF tracking, helping to reduce the cost barrier typically imposed by traditional tracking systems for automated skill assessment.

Dataset

For the purposes of this study, one ophthalmologist and four ophthalmology residents were recruited to perform cataract surgery on a simulated eye (Philips studio artificial eye, Cataract eye basic hard lens PS-OS-001). Participants performed the first four steps of the procedure, including: paracentesis creation, viscoelastic injection, main wound incision, and capsulorhexis formation. Each participant performed five trials, leading to a total of 25 videos. For each trial, we recorded both the monocular microscope video feed and the tracking information from the Intel RealSense camera. The videos ranged in length from 126 to 340 s with an average length of 180 s.

To train the object detection network, the videos from the surgical microscope were split into individual frames, yielding a total of 19145 images. Each of these images was manually annotated with bounding box locations of the visible portions of the same six tools that were affixed with ArUco markers. The annotations were completed in a two-stage process where one individual annotated all images and these labels were then reviewed for correctness by a second, independent reviewer. As the procedure is typically performed by a single surgeon, there were never more than two tools visible at a time in each image in addition to the lens. The annotation process took roughly 100 h to complete the first stage and roughly 20 h to complete the second stage review. Sample images of the annotations are shown in Fig. 3 and the number of annotations for each tool is shown in Fig. 4. Because the lens of the simulated eye was also clearly visible in the surgical microscope video, we also annotated the images with the bounding box location of the lens. The lens appeared in all but 12 images in the total dataset, yielding a total of 19133 annotations.

Object detection network

To obtain tracking information from the surgical microscope video, we trained a YOLOv8 (small) object detection network to recognize the same surgical tools tracked using ArUco. This architecture was chosen for its computational efficiency, strong performance on detection tasks, and accessibility. The network was trained using a leave-one-user-out cross-validation scheme, where data from one participant was reserved for testing, and data from the remaining participants was used for training and validation. This process was repeated for all five participants, resulting in fivefold. Network performance was evaluated using mean average precision at an intersection-over-union threshold of 50% (mAP50). mAP50 is calculated as the mean of the average precision (AP) values for each tool, where AP is the area under the precision–recall curve computed across a range of confidence thresholds at a fixed IoU threshold of 50%. This metric provides a comprehensive evaluation of the model's ability to balance precision and recall for detecting each tool.

Motion-based metrics

Using each tracking modality, we compute a series of motion-based metrics for each video. For each of the six tools used in the procedure, we compute both the path length and the usage time. We selected these metrics as they have been widely shown to correlate with skill for a variety of procedures. For each of the tracking modalities, the calculation of these metrics differs slightly. A full list of all computed metrics is shown in Table 1.

6-DOF tracking

The tool-tip position is obtained from the marker location by performing both pivot and spin calibrations on each tool prior to recording the trials using the pivot calibration module in the SlicerIGT (www.slicerigt.org) extension of 3D Slicer. The calibrations were deemed acceptable when the root mean squared error fell below 1 mm for both pivot and spin calibrations. Given that ArUco tracking is designed to provide 6-DOF navigation, the path length for each tool is defined as the total distance in millimeters traveled by the tool tip. To accurately distinguish between active tool usage and idle periods when tools were placed on the bench but still visible to the RealSense camera, we defined usage time as the time where the magnitude of the change in tool-tip position between consecutive frames exceeded 0.1 mm. This threshold was necessary to align the 6-DOF tracking definition of usage with object detection, where tools are only visible within the field of view of the microscope when actively in use.

Object detection

Unlike ArUco tracking, object detection is not currently capable of providing full 6-DOF measurements of motion. To calculate path length for object detection, we generated predictions using a confidence score threshold of 0.25, a fairly permissive threshold chosen to ensure that potential detections were not prematurely excluded. For each class in each frame, we selected the bounding box with the highest confidence, which eliminates the need for an explicit confidence threshold beyond this step. This approach was appropriate because we know that only one of each instrument is included in the kit, ensuring there is no ambiguity in selecting the correct detection. We then measured the Euclidean distance between matching corners of an instrument's bounding box in consecutive frames. These four corner distances are summed for each frame transition across the entire video and then added together to yield the instrument's total path length. As we do not know the transformation between the image coordinate system and the physical space, we compute the path length in pixels. If a tool is not visible for more than 2 s, the distance traveled between sightings is not included in the path length calculation to avoid inflating the measurement with unobserved motion. Usage time was calculated as the number of occurrences of a tool multiplied by the inverse of the video frame rate.

Lens-based metrics

Maintaining control of the patient's eye movement is a crucial skill in cataract surgery. Surgeons must prevent excessive rotations of the eye to ensure that the lens remains visible and centered within the surgical microscope's field of view. While conventional marker-based tracking methods cannot measure eye movement, the lens is clearly visible in the microscope video and has a fixed location relative to the rest of the eye. To assess eye movement, we defined several lens-based metrics that could be computed using object detection.

Motion-based metrics of the lens: Similar to the motion-based metrics calculated for each tool used in the procedure, we computed the path length and usage time of the lens as previously defined for the tools in "Object detection".

Distance from center: To evaluate how well the surgeon kept the eye in the center of the field of view, we calculated both the mean and maximum distances from the center of the lens bounding box to the center of the image captured by the surgical microscope video.

Expert scoring

To obtain a ground truth skill evaluation, two expert ophthalmologists were asked to review each video and provide

Table 1 Summary of metrics used for skill predictions

Tool/ Object	Metric	Tracking	Object detection
Procedure	Total time	X	X
Forceps	Path length	X	X
	Usage time	X	X
Diamond keratome (straight)	Path length	X	X
	Usage time	X	X
Viscoelastic cannula	Path length	X	X
	Usage time	X	X
Cystotome needle	Path length	X	X
	Usage time	X	X
Diamond keratome (angled)	Path length	X	X
	Usage time	X	X
Capsulorhexis forceps	Path length	X	X
	Usage time	X	X
Lens	Path length		X
	Usage time		X
	Avg. distance from center		X
	Max. distance from center		X

Scoring Criteria

Task-specific assessment:

(4) Corneal entry
(5) Paracentesis and viscoelastic entry
(6) Capsulorhexis: Commencement of flap and follow-through
(7) Capsulorhexis: Formation and circular completion

Global assessment:

(14) Wound neutrality and minimizing eye rolling and corneal distortion
(15) Eye positioned centrally within microscope view
(16) Conjunctival and corneal tissue handling
(17) Intraocular spatial awareness

Scoring assignment: 2 = Novice, 3 = Beginner, 4 = Advanced Beginner, 5 = Competent

Fig. 5 Scoring criteria from ICO-OSCAR: SICS rubric used for expert evaluation

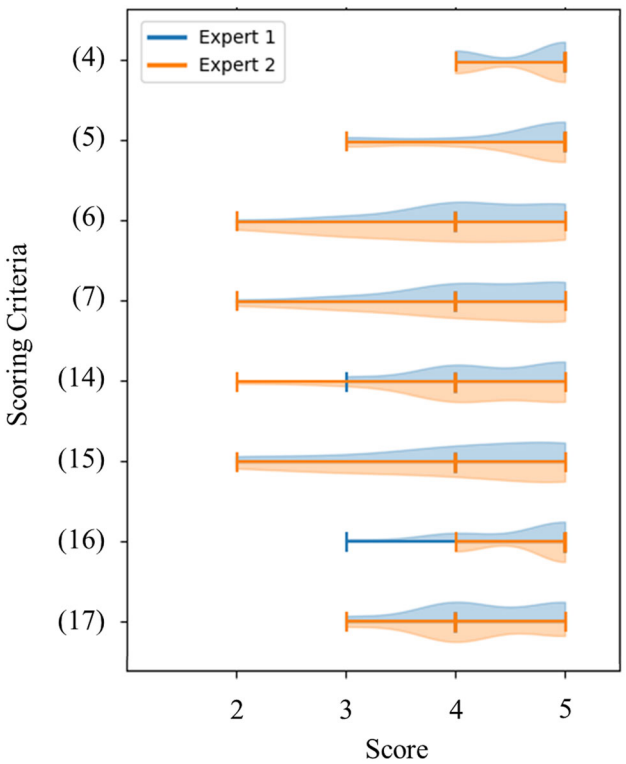

Fig. 6 Distribution of expert-assigned scores for each scoring criterion of ICO-OSCAR:SICS rubric. The hash-marks indicate the minimum, median, and maximum scores provided by the experts

scores using the International Council of Ophthalmology's ophthalmology surgical assessment rubric for small incision cataract surgery (ICO-OSCAR:SICS) rubric, which is widely used for assessing cataract surgery [20]. This rubric involves assigning a score from two to five for each criterion, where five indicates expert proficiency and two indicates novice performance. This scale separates skill evaluation into both a task-specific assessment where the surgeon is evaluated on how well they performed individual steps in the procedure, and a global assessment where they are evaluated on criteria related to their overall performance. In cases where the two experts' scores differed, we computed the average score and rounded up to the nearest whole number to assign the final score for that criterion. This is the same rubric that was used to assess trainees in [19]; however, the authors in that study examined only the task-specific criteria for a single step of the procedure. For this study, participants performed the first four tasks in the procedure, and so we evaluate predictions on the first four task-specific criteria as well as four global assessment criteria. The rating scale used is shown in Fig. 5. The full distribution of the expert-assigned scores for each

criterion is shown in Fig. 6. To assess the inter-rater reliability, we computed the intraclass correlation coefficient (ICC) between the scores provided by the two expert reviewers. The ICC between the two reviewers was 0.96.

Experiments

Correlation of raw metric values with expert scores

The objective of this experiment is to determine whether the computed metrics from both 6-DOF tracking and object detection hold predictive value for skill assessment in cataract surgery. To achieve this objective, we directly compare each raw metric to each expert-assigned scoring criterion using Spearman rank correlation. This approach allows us to assess the strength and direction of associations between the individual metrics generated by each tracking modality (6-DOF and object detection) and the expert scores.

Predicting expert scores using ordinal regression

In this experiment, we aim to assess the effectiveness of metrics from both 6-DOF tracking and object detection in predicting expert-assigned ICO-OSCAR:SICS scores. For each scoring criterion, we selected up to five metrics with the strongest, statistically significant correlations with the ground truth scores based on the results of the previous experiment. If no metrics had significant correlations, a single metric with the highest, nonsignificant correlation was chosen.

Using these selected metrics, we trained an ordinal regression model to predict the ICO-OSCAR:SICS scores. The model was trained using the same cross-validation strategy applied in training the object detection network. Predictions from the model were then compared to the ground truth scores using Spearman rank correlation. This metric was chosen because it assesses the strength and direction of monotonic relationships, which aligns with the ordinal nature of the ICO-OSCAR:SICS scores. Additionally, its use has been well established in several studies evaluating the similarity between predicted and expert-assigned scores in surgical performance assessment [21, 22].

To evaluate differences in correlation strengths, we compared 6-DOF tracking to both object detection experiments (with and without lens-based metrics). For this comparison, correlation values were first converted to z-scores using the Fisher transformation and then analyzed using a z-test.

This experiment was run twice: first, using only metrics computable by both 6-DOF tracking and object detection, and second, with lens-based metrics included. In the second analysis, we retained the same metrics selected for each criterion from the first analysis and added the lens-based metric with the strongest significant correlation with the ground truth scores (up to one additional metric per criterion). If no lens-based metrics had significant correlations, no additional metrics were included, as the metrics selected in the first analysis were sufficient to make predictions. This approach ensured that lens-based metrics were treated as supplementary information rather than a requirement.

When analyzing the performance of the score predictions, we considered negative correlations to be poor predictions of skill because the expert scores and predicted scores should align in value and direction. Positive correlations indicate that the predicted scores track the expert-assigned scores accurately, reflecting agreement in skill assessment. As in the initial analysis, predicted scores were compared to ground truth scores using Spearman rank correlation.

Table 2 mAP50 results for object detection network

Tool / Object	mAP50
Forceps	73.4%
Diamond keratome (straight)	69.0%
Viscoelastic cannula	60.1%
Cystotome needle	61.5%
Diamond keratome (angled)	66.0%
Capsulorhexis forceps	78.3%
Lens	98.0%
Overall	72.6%

Results

Object detection network performance

The mAP50 for each of the tools and overall is presented in Table 2. Using the YOLOv8 object detection network, we achieved an overall mAP50 of 75.6%. The lens was detected with a mAP50 of 98.0%. The tool with the highest mAP50 was the capsulorhexis forceps at 78.3%, while the tool with the lowest mAP50 was the viscoelastic cannula at 60.1%.

Correlation of raw metric values with expert scores

In the analysis of correlation strengths between each metric and the expert scores, object detection metrics generally demonstrated stronger correlations with the scoring criteria than the 6-DOF tracking metrics. When comparing the number of significant correlations, object detection metrics showed a higher frequency of significant associations with expert scores. Specifically, for 6-DOF tracking, there were 14 instances of significant correlations with expert scores out of a maximum of 104 comparisons, with at least one signifi-

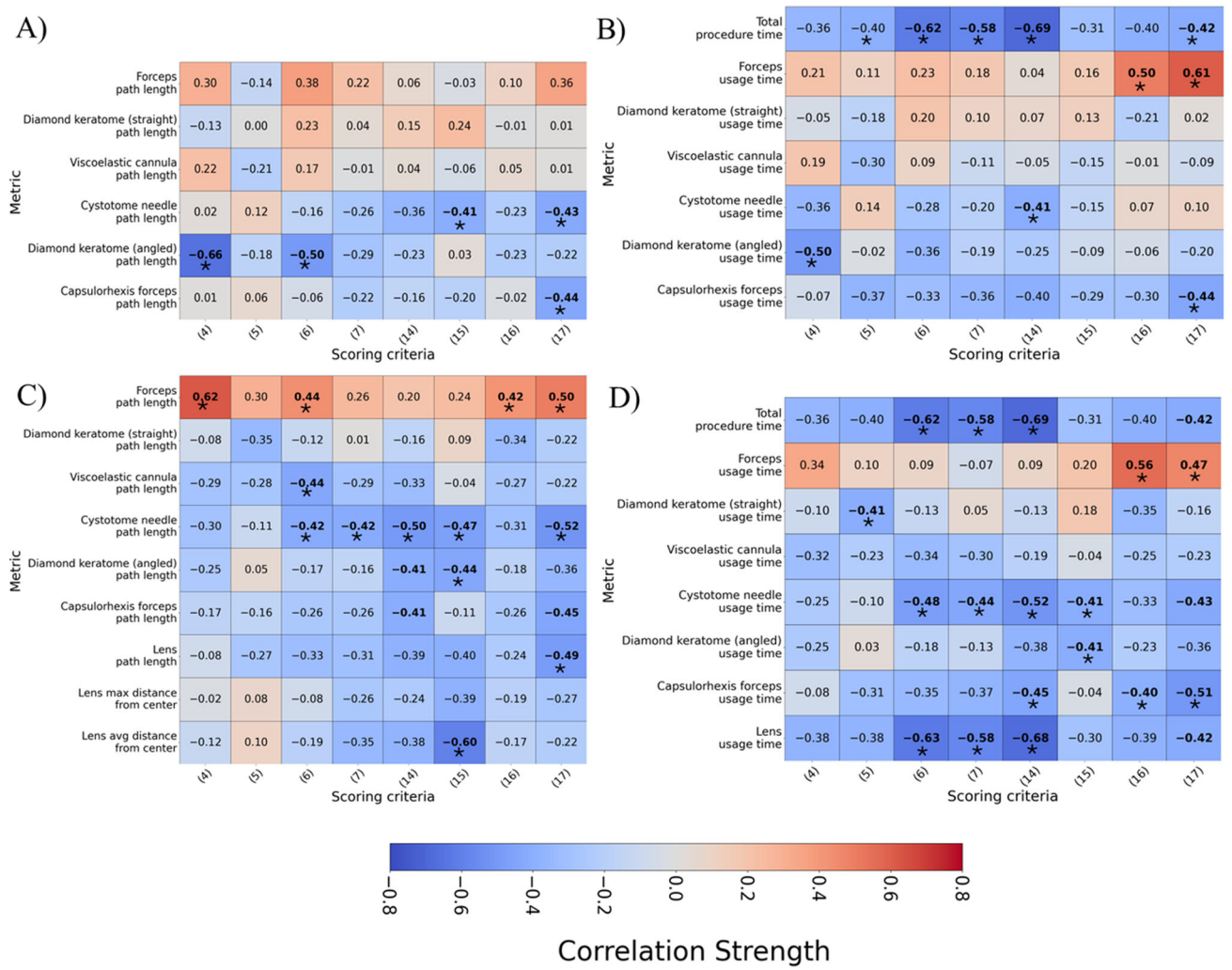

Fig. 7 Spearman rank correlations between metric values and expert-assigned scoring criteria. Subplots **A** and **B** display the correlation strengths for distance-based and time-based metrics, respectively, derived from 6-DOF tracking. Subplots **C** and **D** show the corresponding correlations for distance-based and time-based metrics computed from object detection. Positive correlations are indicated in red, and negative correlations in blue, with significant correlations highlighted in bold. Metrics that are used for score prediction are denoted with an asterisk (*)

cant correlation for seven out of the eight scoring criteria. In contrast, object detection metrics yielded 30 significant correlations out of 104 comparisons without lens-based metrics and 36 significant correlations out of 144 comparisons when lens-based metrics were included, with at least one significant correlation for each of the eight scoring criteria.

Across both 6-DOF tracking and object detection, several metrics exhibited negative correlations with the expert scores, suggesting that higher scores corresponded to reduced metric values, which may imply greater efficiency in movements. The only metric category that predominantly showed positive correlations with the expert scores was the forceps, which consistently demonstrated positive associations in both 6-DOF tracking and object detection results. The complete results of the correlations between metrics and expert scores are shown in Fig. 7.

Predicting expert scores using ordinal regression

In this experiment, we evaluated the correlations between predicted scores and expert-assigned scores across three conditions: scores predicted using metrics from 6-DOF tracking, object detection without lens-based metrics, and object detection with lens-based metrics. The correlations between the predicted scores and expert scores are shown in Table 3.

When using object detection metrics without lens-based metrics, predicted scores showed stronger correlations with expert scores than those from 6-DOF tracking for four out

Table 3 Spearman rank correlations between predicted scores and expert scores for each scoring criterion

Scoring criteria	6-DOF tracking	Object detection **(no lens metrics)**	Object detection **(with lens metrics)**
Task-specific assessment			
(4) Corneal Entry	−0.15	0.16	**0.48**
	($p = 0.46$)	($p = 0.43$)	(**p = 0.02**)
(5) Paracentesis and Viscoelastic Entry	0.37	0.18	**0.80**
	($p = 0.06$)	($p = 0.38$)	(**p < 0.001**)
(6) Capsulorrhexis: Commencement of Flap and Follow-Through	0.19	**0.55**	**0.52**
	($p = 0.38$)	(**p = 0.004**)	(**p = 0.007**)
(7) Capsulorrhexis: Formation and Circular Completion	0.19	0.26	0.28
	($p = 0.37$)	($p = 0.22$)	($p = 0.17$)
Global assessment			
(14) Wound Neutrality and Minimizing Eye Rolling and Corneal Distortion	0.23	0.10	0.31
	($p = 0.26$)	($p = 0.62$)	($p = 0.12$)
(15) Eye Positioned Centrally Within Microscope View	0.37	0.23	**0.42**
	($p = 0.07$)	($p = 0.25$)	(**p = 0.04**)
(16) Conjunctival and Corneal Tissue Handling	**0.61**	**0.72**	**0.72**
	(**p = 0.001**)	(**p < 0.001**)	(**p < 0.001**)
(17) Intraocular Spatial Awareness	−0.40	**0.51**	**0.55**
	($p = 0.05$)	(**p = 0.01**)	(**p = 0.005**)

Significant correlations ($p < 0.05$) are highlighted in bold

of the eight scoring criteria. The addition of lens-based metrics further increased correlation strengths across all eight criteria, resulting in six criteria with significant positive correlations between predicted scores and expert scores.

The strongest correlation for 6-DOF tracking was observed for criterion (16): Conjunctival and Corneal Tissue Handling ($\rho = 0.61$, $p = 0.001$), while the least predictive correlation was for criterion (17): Intraocular Spatial Awareness ($\rho = -0.40$, $p = 0.05$), as the negative direction indicates limited utility for skill prediction. For object detection without lens-based metrics, criterion (16): Conjunctival and Corneal Tissue Handling also displayed the highest correlation ($\rho = 0.72$, $p = 5.4 \times 10^{-5}$), whereas criterion (14): Wound Neutrality and Minimizing Eye Rolling and Corneal Distortion had the lowest ($\rho = 0.10$, $p = 0.62$). When lens-based metrics were included, criterion (5): Paracentesis and Viscoelastic Entry exhibited the strongest correlation ($\rho = 0.80$, $p = 1.4 \times 10^{-6}$), while criterion (7): Capsulorrhexis: Formation and Circular Completion had the weakest ($\rho = 0.28$, $p = 0.17$). Only one criterion, (6): Capsulorrhexis: Commencement of Flap and Follow-Through, showed a decrease in correlation when lens-based metrics were added, indicating that additional metrics did not improve prediction accuracy across all criteria.

Across all conditions, there were no criteria where 6-DOF tracking produced significantly higher correlations with expert scores compared to object detection ($p > 0.07$). These results indicate that, for all scoring criteria, object detection methods performed at least as well as, and often better than, 6-DOF tracking in terms of alignment with expert scores. The $p > 0.07$ value represents the lowest p value obtained for any scoring criterion when comparing the three methods (6-DOF tracking, object detection without lens-based metrics, and object detection with lens-based metrics). Conversely, object detection, without lens-based metrics, produced significantly stronger correlations for criteria (4): Corneal Entry, (7): Capsulorrhexis: Formation and Circular Completion, and (17): Intraocular Spatial Awareness, with p values of 0.01, 7.8×10^{-4}, and 6.6×10^{-13}, respectively. With the inclusion of lens-based metrics, these differences became more pronounced for these criteria, and criterion (5): Paracentesis and Viscoelastic Entry also showed a significantly stronger correlation ($p = 1.4 \times 10^{-6}$) compared to the correlation from 6-DOF tracking.

Discussion

The results of this study demonstrate that object detection from monocular surgical video can achieve comparable, and in some cases stronger, correlations with expert-assigned skill scores than traditional 6-DOF ArUco-based tracking. When comparing the raw metrics to expert scores, object detection metrics exhibited more frequent and generally stronger correlations across multiple scoring criteria compared to 6-DOF tracking, particularly when lens-based metrics were included. These findings suggest that object detection could be a valuable tool for skill assessment, especially given its accessibility and lack of additional hardware requirements. Furthermore, object detection analyzes motion directly from the surgical microscope's viewpoint, which closely aligns with the perspective used by expert raters to evaluate performance. This alignment not only ensures that the assessment framework reflects real-world evaluation practices but also emphasizes the practical relevance of video-based analysis in skill assessment. In the score prediction analysis, correlations between the predicted scores and expert scores were consistently higher for object detection metrics than for 6-DOF tracking, with lens-based metrics further enhancing prediction accuracy. These results support the hypothesis that object detection can provide a robust assessment of surgical skill and may offer advantages over tracking methods that require attachments to the surgical tools, especially when predictive power is a key consideration.

The higher correlations observed with object detection, especially when using lens-based metrics, indicate that video-based analysis can capture nuanced aspects of instrument handling and movement that may be challenging to detect with optical 6-DOF tracking methods under certain conditions, such as occlusion or challenging orientations. As with any optical method, ArUco tracking can suffer from line-of-sight issues, leading to frames being missed when instruments are occluded by other objects. This issue occurs much less frequently in the microscope video because it is optimally positioned to ensure the surgeon can see their work. While sensor-based methods such as electromagnetic (EM) tracking could address line-of-sight issues, they introduce their own challenges, including interference from metallic objects, calibration complexity, and the need for additional hardware attached to the instruments. These limitations underscore the potential of object detection to complement or replace existing tracking methods, providing a robust, scalable, and accessible alternative for assessing surgical skill.

Interestingly, the observed negative correlations for some metrics suggest that higher skill levels, as perceived by experts, are often associated with more efficient and conservative instrument movements. This pattern was consistently captured by both object detection and 6-DOF tracking, highlighting the value of both methods for skill assessment. Additionally, forceps usage tended to show positive correlations, indicating that expert evaluators may associate this instrument's usage with more precise and targeted actions. Unlike other instruments, the opening and closing motion of the forceps contributes to path length, which may explain why its usage differs in correlation patterns compared to other

tools. This distinction suggests that tool-specific movement characteristics could provide additional insight into surgical technique, reinforcing the potential for object detection to capture nuanced aspects of performance.

To enable practical implementation, we envision integrating this method as a feature within the existing Perk Tutor platform in 3D Slicer. This approach would make the system freely available as open-source software, facilitating easy integration with commonly available hardware. Our ultimate goal is to provide continuous feedback on surgical performance without requiring the presence of an expert, thereby streamlining the assessment process and reducing the burden on teaching physicians. Additionally, by providing objective, automated feedback aligned with OSCAR scoring criteria, this approach could serve as a formative tool to help students identify specific areas for improvement in their surgical technique. Such an approach would enable frequent, detailed feedback to support skill development, complementing the role of the teaching physician in guiding the student's progress.

Despite these promising results, this study has several limitations. First, the small dataset size (25 videos) constrained the range of expert scores and limited the statistical power to draw definitive conclusions about score classification accuracy. The limited data also necessitated a simpler approach to metric selection, where only the top correlating metrics were used to predict scores. Additionally, the small dataset size made it impractical to use cross-validation for metric selection. While the ordinal regression models were still trained using cross-validation, the inability to perform cross-validation for metric selection further restricts our ability to make strong conclusions about the accuracy of the predicted scores. Furthermore, for some criteria, such as Criterion 16 (Conjunctival and Corneal Tissue Handling), the limited range of scores made predictions easier, as binary outputs are inherently simpler to model than criteria with wider score distributions. This may partially explain why Criterion 16 showed strong predictive performance. Nevertheless, the comparison between tracking modalities remains valid. Although the results support the hypothesis that object detection can offer a viable alternative to 6-DOF tracking, future studies with larger and more diverse datasets will be necessary to validate these findings and potentially apply more sophisticated metric selection techniques or machine learning classifiers for more accurate score predictions.

Building on this comparison, it is important to note that although 6-DOF tracking provides both translational and rotational data, this study focused exclusively on metrics derived from tool-tip movement, such as path length and usage time. This decision was driven by the need to ensure a fair comparison with object detection, which does not currently provide rotational or pose information. Additionally, tool-tip metrics have been widely validated in the literature as reliable indicators of surgical skill. However, excluding rotational and pose-based metrics represents a limitation, as such data could offer valuable additional insights into surgical performance. Future research could explore the integration of rotational metrics to further enhance the predictive power of 6-DOF tracking and assess their potential contribution to skill evaluation.

Another limitation is that only ArUco markers were used for 6-DOF tracking. While ArUco is a low-cost, accessible option and is widely considered a standard in marker-based optical tracking, it is known to have limitations in accuracy, particularly in complex orientations or when markers are partially occluded. Additionally, the recording frame rate of 6 fps, constrained by the limitations of the ArUco tracking software and the laptop used for recording, is lower than typical video frame rates. However, the graders reported no difficulty evaluating the videos at this frame rate, suggesting that the key features for assessment were adequately preserved. These factors could introduce bias into the comparisons between ArUco and object detection, potentially underestimating the performance of 6-DOF tracking. Future work could address these limitations by validating object detection against higher-precision, high-cost tracking systems to more comprehensively assess its robustness relative to premium tracking modalities and by exploring the impact of higher frame rates on assessment fidelity.

Another important consideration is the use of Spearman correlation to assess agreement between predicted and expert-assigned scores. While Spearman correlation is widely used in skill assessment research, it has inherent limitations. Because it evaluates only rank order rather than absolute differences, it does not capture the magnitude of prediction errors. This means that large deviations in score values may still yield high correlations if the ranking order is preserved. Additionally, Spearman correlation is less informative when many tied values exist, as is often the case with ordinal skill assessment scores. To provide a more comprehensive evaluation of predictive performance, future work could explore complementary metrics such as mean absolute error (MAE) or quadratic weighted kappa (QWK), which account for absolute differences and agreement levels across score categories. Incorporating these additional metrics could offer deeper insights into prediction accuracy and the practical utility of automated skill assessment methods.

The process of manual annotation, while critical for the success of this study, posed a significant time burden, requiring approximately 120 h to annotate and review 1.25 h of video. Although this high time cost is challenging, annotated data is essential for developing and validating new methodologies, such as object detection for skill assessment. Unlike calibration for 6-DOF tracking, which must be performed each time the system is used to ensure proper functioning, annotation requires time only from those

developing the system and can be done asynchronously as more data becomes available. Additionally, as the object detection model improves, the need for further annotation diminishes, whereas calibration must be repeated regardless of system improvements. The availability of high-quality, human-annotated datasets ensures reproducibility, rigor, and a foundation for training automated systems. Future research could explore semi-automated or AI-assisted annotation tools to accelerate this process, thereby reducing the time investment required while maintaining data quality.

Finally, while object detection has significant advantages in terms of cost, hardware requirements, and ease of integration, real-world clinical environments present additional challenges, such as varying lighting conditions, reflections, and diverse tool types, which will reduce the robustness of video-based tracking. Despite these challenges, the object detection model used in this study allowed us to compute metrics that effectively distinguished between different levels of surgical performance, suggesting that the model's performance is sufficient for this task. Nonetheless, there remains room for improvement in the model's accuracy and robustness. Future studies should aim to enhance the model to ensure consistent performance in diverse clinical scenarios, improving the resilience of object detection models in dynamic, real-world surgical environments.

Conclusion

This study demonstrates that object detection from monocular surgical video can effectively assess surgical skill in cataract procedures, achieving comparable or stronger correlations with expert-assigned scores than 6-DOF ArUco tracking. By eliminating the need for additional hardware or tool modifications, object detection offers a scalable and accessible solution for skill evaluation. The inclusion of lens-based metrics further enhanced predictive accuracy, underscoring the potential of video-based analysis to provide meaningful insights into surgical performance.

Despite challenges such as manual annotation and variability in clinical environments, the metrics derived from object detection reliably distinguished between different levels of surgical skill, demonstrating the model's suitability for this application. Future studies should aim to validate this approach with larger and more diverse datasets, improve its robustness to real-world variability, and explore its integration into clinical and training settings as a practical tool for skill assessment.

Acknowledgements This work is funded in part by Natural Sciences and Engineering Research Council of Canada Grant RGPIN-2020-05582, by the NSERC Discovery Grant program RGPIN-2022-03919, and by the Southeastern Ontario Academic Medical Organization (SEAMO) Endowed Scholarship and Education Fund. R. Hisey is supported by the NSERC Canada Graduate Scholarship (Doctoral). G. Fichtinger is supported as an NSERC Canada Research Chair in Computer-Integrated Surgery.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethics approval Approval was obtained from the ethics committee of Queen's University. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Informed Consent Informed consent was obtained from all individual participants included in the study.

References

1. Cremers SL, Ciolino JB, Ferrufino-Ponce ZK, Henderson BA (2005) Objective assessment of skills in intraocular surgery (OASIS). Ophthalmology 112(7):1236–1241
2. Dean W (2018) Competency training: using the ico cataract rubric to learn and teach cataract surgery. Commun Eye Health J 31(102):44–44
3. Gauba V, Tsangaris P, Tossounis C, Mitra A, McLean C, Saleh GM (2008) Human reliability analysis of cataract surgery. Arch Ophthalmol 126(2):173–177
4. Eaton E, Tarpley JL, Solorzano CC, Cho CS, Weber SM, Termuhlen PM (2011) Resident education in 2011: three key challenges on the road ahead. Surgery 149(4):465–473
5. Williams RG, Verhulst S, Colliver JA, Dunnington GL (2005) Assuring the reliability of resident performance appraisals: more items or more observations? Surgery 137(2):140–147
6. Ott MC, Pack R, Cristancho S, Chin M, Koughnett J, Ott M (2022) "the most crushing thing": understanding resident assessment burden in a competency-based curriculum. J Grad Med Educ 14(5):583–592
7. Tierney AA, Rosner BI (2023) Clinical assessment of residents: a survey of clinician educators regarding resident assessment burden and modifiable factors. J Grad Med Educ 15(1):92–97
8. Polin MR, Siddiqui NY, Comstock BA, Hesham H, Brown C, Lendvay TS, Martino MA (2016) Crowdsourcing: a valid alternative to expert evaluation of robotic surgery skills. Am J Obstet Gynecol 215(5):644–16447
9. Clinkard D, Holden M, Ungi T, Messenger D, Davison C, Fichtinger G, McGraw R (2015) The development and validation of hand motion analysis to evaluate competency in central line catheterization. Acad Emerg Med 22(2):212–218
10. Saleh GM, Gauba V, Sim D, Lindfield D, Borhani M, Ghoussayni S (2008) Motion analysis as a tool for the evaluation of oculoplastic surgical skill: evaluation of oculoplastic surgical skill. Arch Ophthalmol 126(2):213–216
11. Stenmark M, Omerbašić E, Magnusson M, Andersson V, Abrahamsson M, Tran PK (2022) Vision-based tracking of surgical motion during live open-heart surgery. J Surg Res 271:106–116
12. Sharon Y, Jarc AM, Lendvay TS, Nisky I (2021) Rate of orientation change as a new metric for robot-assisted and open surgical skill evaluation. IEEE Trans Med Robot Bionics 3(2):414–425
13. Poursartip B, LeBel ME, Patel RV, Naish MD, Trejos AL (2018) Analysis of energy-based metrics for laparoscopic skills assessment. IEEE Trans Biomed Eng 65(7):1532–1542

14. Trejos AL, Patel RV, Malthaner RA, Schlachta CM (2014) Development of force-based metrics for skills assessment in minimally invasive surgery. Surg Endosc 28:2106–2119
15. Brown JD, O'Brien CE, Leung SC, Dumon KR, Lee DI, Kuchenbecker KJ (2017) Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer. IEEE Trans Biomed Eng 64(9):2263–2275
16. Choi H, Park Y, Lee S, Ha H, Kim S, Cho HS, Hong J (2017) A portable surgical navigation device to display resection planes for bone tumor surgery. Mini Invasive Ther Allied Technol 26(3):144–150
17. O'Driscoll O, Hisey R, Camire D, Erb J, Howes D, Fichtinger G, Ungi T (2021) Object detection to compute performance metrics for skill assessment in central venous catheterization. Proc. SPIE Medical Imaging 2021: Image-Guided Procedures. Robot Interven Model 11598:315–322
18. Chan RW, Hisey R, Holden MS (2021) Feasibility of video-based skills assessment: a study on ultrasound-guided needle insertions using simulated projections. Proc. SPIE Medical Imaging 2021: Image-Guided Procedures. Robot Interven Model 12034:663–669
19. Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS (2019) Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int J Comput Assist Radiol Surg 14(6):1097–1105
20. Golnik KC, Beaver H, Gauba V, Lee AG, Mayorga E, Palis G, Saleh GM (2011) Cataract surgical skill assessment. Ophthalmology 118(2):427
21. Benmansour M, Malti A, Jannin P (2023) Deep neural network architecture for automated soft surgical skills evaluation using objective structured assessment of technical skills criteria. Int J Comput Assist Radiol Surg 18(5):929–937
22. Kelly JD, Petersen A, Lendvay TS, Kowalewski TM (2020) Bidirectional long short-term memory for surgical skill classification of temporally segmented tasks. Int J Comput Assist Radiol Surg 15(12):2079–2088