Computers in Biology and Medicine 141 (2022) 105121

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Automated tool detection with deep learning for monitoring kinematics and eye-hand coordination in microsurgery

Jani Koskinen [a,*], Mastaneh Torkamani-Azar [a], Ahmed Hussein [b,d], Antti Huotarinen [b,c], Roman Bednarik [a]

[a] School of Computing, University of Eastern Finland, Länsikatu 15, Joensuu, 80100, Pohjois-Karjala, Finland
[b] Microsurgery Center, Kuopio University Hospital, Kuopio, 70211, Pohjois-Savo, Finland
[c] Department of Neurosurgery, Institute of Clinical Medicine, Kuopio University Hospital, Kuopio, 70211, Pohjois-Savo, Finland
[d] Department of Neurosurgery, Faculty of Medicine, Assiut University, Assiut, 71111, Egypt

ARTICLE INFO

Keywords:
Microsurgery
Tool detection
Tool kinematics
Eye tracking
Eye-hand coordination
Computer vision
Transfer learning
Convolutional neural networks

ABSTRACT

In microsurgical procedures, surgeons use micro-instruments under high magnifications to handle delicate tissues. These procedures require highly skilled attentional and motor control for planning and implementing eye-hand coordination strategies. Eye-hand coordination in surgery has mostly been studied in open, laparoscopic, and robot-assisted surgeries, as there are no available tools to perform automatic tool detection in microsurgery. We introduce and investigate a method for simultaneous detection and processing of micro-instruments and gaze during microsurgery. We train and evaluate a convolutional neural network for detecting 17 microsurgical tools with a dataset of 7500 frames from 20 videos of simulated and real surgical procedures. Model evaluations result in mean average precision at the 0.5 threshold of 89.5–91.4% for validation and 69.7–73.2% for testing over partially unseen surgical settings, and the average inference time of 39.90 ± 1.2 frames/second. While prior research has mostly evaluated surgical tool detection on homogeneous datasets with limited number of tools, we demonstrate the feasibility of transfer learning, and conclude that detectors that generalize reliably to new settings require data from several different surgical procedures. In a case study, we apply the detector with a microscope eye tracker to investigate tool use and eye-hand coordination during an intracranial vessel dissection task. The results show that tool kinematics differentiate microsurgical actions. The gaze-to-microscissors distances are also smaller during dissection than other actions when the surgeon has more space to maneuver. The presented detection pipeline provides the clinical and research communities with a valuable resource for automatic content extraction and objective skill assessment in various microsurgical environments.

1. Introduction

Eye and hand movements are closely connected in a range of tasks. When reaching for an object, humans direct their gaze towards the object [5,30]. Eye movements also correlate with motor control learning and decision making processes [4,13,15,20] — associations that have motivated new applications like gaze training [5] and that make eye-hand coordination a central skill in manual tasks that require effective aiming and motor strategies [10,31,41,55]. Thus, it is important to understand how eye-hand coordination develops, how it exhibits in experts, and how it is affected by ergonomics and environmental factors.

Studies on eye-hand coordination in surgery have mostly focused on general [19], endoscopic [53], laparoscopic [58], and robot-assisted surgeries [63]. Hardly any studies have investigated eye-hand coordination in challenging microsurgical procedures where surgical microscopes provide highly magnified views of tissues and micros instruments.

1.1. Methods for eye-hand coordination analysis in microsurgery

With the availability of reliable sensors for operating rooms, eye tracking methodology has been applied to gain an understanding of several aspects of surgical procedures. For example, in laparoscopy training, Law et al. applied eye tracking to learn how to differentiate gaze patterns of expert and novice surgeons [32], Khan et al. modeled

* Corresponding author.
E-mail address: jani.koskinen@uef.fi (J. Koskinen).

https://doi.org/10.1016/j.compbiomed.2021.105121
Received 14 September 2021; Received in revised form 30 November 2021; Accepted 3 December 2021
Available online 11 December 2021

differences in gaze patterns of novice and expert surgeons in real laparoscopy procedures [27], and Tien et al. used remote eye tracking to show that gaze patterns can be cued to more effective locations in laparoscopic training [54]. In robotic surgery, Leff et al. implemented expert visual guidance in the form of gaze tracking to improve performance and reduce attentional demands [33]. Nevertheless, due to the limited availability of suitable eye tracking technologies [12], only a few studies have examined eye-gaze related metrics in microsurgery [2,6,13,29].

The joint analysis of gaze and tool movements has been shown to outperform single modalities in surgical activity recognition [53]. Microsurgical procedures, however, are tremendously challenging not only because they demand finely tuned eye-hand coordination [17], but also because of the complexity of tracking tool movements. In open surgery, tools can be tracked using auxiliary instrumentation such as infrared markers attached to the tools, and cameras strategically aligned with the surgical field to capture the marker movements [49]. These systems are not feasible in many microsurgical procedures where the surgeon's hands make minimal movements and the available space is limited [3]. However, surgical microscopes provide high-quality video recordings that can be used for instrument detection through modern computer vision solutions [11,34].

1.2. Computer vision methods for tool detection in surgery

Surgical tools' motions and presence in the task can be used to evaluate technical performance [49] and classify phases of surgical procedures [18,51]. As these are central aims for several envisioned computer assisted surgical systems, researchers have investigated methods for detecting the surgical tools automatically [38,50,60]. Some methods have focused on a single instrument, such as surgical needles [40], but most have detected multiple tools [see, for example [18,25,34,45,59]].

Previous studies have relied on models purposely built for surgical tool detection that typically use convolutional neural network (CNN) architectures for bounding box detection or area segmentation [8,60]. Others have extended these approaches by using recurrent neural networks (RNN) and long short-term memory (LSTM) models that use information from multiple frames for making predictions [45]. Some researchers have evaluated or extended existing open-source models for surgical tool detection. For example, some studies detected laparoscopic instruments using the "You Only Look Once" (YOLO) [48] detector [7,59], including its modified version that also utilizes temporal information [25].

1.3. Technical contributions

We introduce a novel method for analysing bi-manual actions jointly with gaze during microsurgery. The method utilizes an open-source single-shot detector – YOLOv5l from the YOLOv5 repository [26] – whose main advantage is its speed that enables the online analysis of tool movements concurrently with other modalities, such as electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), and gaze tracking [33]. The technical contributions of this study are:

1. Yolov5l is trained and evaluated for automatic detection of 17 micro-instruments from videos of real and simulated microsurgical procedures. The robustness of the model is reported through its high accuracy and fast inference from a dataset of 20 distinct surgical settings, thus surpassing previous studies that focused on surgical tool detection from homogeneous datasets [11,59].
2. We evaluate how patterns learned from simulated training tasks can be transferred to tool detection in real microsurgical operations. Three experiments are conducted to test (1) how the detector performs in partially unseen environments; (2) how the performance improves if the training is continued with a small amount of data from unseen environments; and (3) how the detector performs on videos of real surgical procedures if training data from real surgical procedures is supplemented with data from simulated procedures.
3. A successful application of the model is presented by combining it with a state-of-the-art microscope eye-tracker to investigate surgeon's eye-hand coordination during simulated intracranial vessel dissection. A pipeline is introduced for processing and analysing tool kinematics in order to characterize the surgeon's bi-manual actions. To the best of our knowledge, this is the first approach to evaluating microsurgeon's eye-hand coordination directly through eye and hand movements. The system's performance is evaluated under two different surgical approaches, and we identify potential limitations that have not been assessed in previous studies.

The rest of this paper is split into two main sections. Section 2 describes the microsurgical tool detection model, datasets, and experiments. Section 3 is dedicated to the case study where the tool and eye tracking system was tested by analyzing tool kinematics and gaze-tool interactions in two different microsurgical approaches. Finally, discussions of detection accuracy and inference time, interpretations of tool kinematics and eye-hand coordination outcomes, and suggestions for future work are presented in Section 4.

2. Microsurgical tool detection: a deep learning approach

In this section, we introduce a computer vision pipeline for microsurgical tool detection in intraoperative and training scenarios.

2.1. Network architecture and evaluation criteria

The proposed pipeline for surgical tool detection is based on YOLOv5-l, the large version of YOLOv5 released in June 2020. YOLOv5-l is based on CNNs with a backbone composed of Cross Stage Partial Network (CSPNet) [57] for feature extraction, and spatial pyramid pooling [23]. The model outputs are the bounding box coordinates, generated at the minimum confidence threshold of 0.001, and detection confidence values for each detected object. Multiple detections of the same object are removed using non-maximum suppression (NMS) with the IoU threshold of 0.6. The network is evaluated using the per class measures of precision, defined as the ratio of true positives to the total number of predictions, and recall, defined as the ratio of true positives to the total number of predictions. For each class, YOLOv5 also calculates the Average Precision (AP), defined as the precision averaged across all recall values between 0 and 1, the F1-score, defined as the harmonic mean of precision and recall, and the Intersection over Union (IoU), the ratio of the intersection and the union of the ground-truth and the detected bounding box [24]. The fitness function that determines the best model is defined as the weighted sum of precision, recall, mean Average Precision (mAP) at the IoU of 0.5 (mAP@0.5), and mAP at the IoU of 0.5–0.95 (mAP@[0.5:0.95]). The default weights for these metrics are 0.0, 0.0, 0.1 and 0.9, respectively.

2.2. Microsurgical dataset preparation

The datasets consisted of video frames collected from multiple sources: (1) videos of participants completing different microsurgical simulation training tasks in a microsurgery training center, and (2) videos of real microsurgical procedures uploaded on public video repositories such as YouTube under the Creative Commons (CC) license. Table 1 describes the microsurgical procedures or training tasks depicted in the videos, and Fig. 1 demonstrates sample frames from these videos. The videos of OR procedures were heavily edited, typically showing a single surgical maneuver or procedure phase for less than 20 s.

The videos include approximately 20 distinct settings as, for example, some of the training tasks used different training boards or a

Table 1
Overview of microsurgery video sources used for tool detection.

	Task	Duration [min]	Resolution
Simulation training tasks	Continuous suturing	20	1280 × 720
	Knot tying/interrupted suturing [2]	240	720 × 486
	Forced accuracy needle piercing/cutting	30	1280 × 720
	Vessel dissection	40	1280 × 720
	Porcine drilling	60	1280 × 720
OR procedures	Various neurosurgical procedures [n = 10], midline suboccipital craniotomy [21], resection of a cavernous malformation [35,39], resection of a IV ventricular subependymoma [37], resection of a cervicomedullary ganglioma [36], two cerebral bypasses [9], resection of a brain glioma [42–44]	Approx. 80 in total	1280 × 720

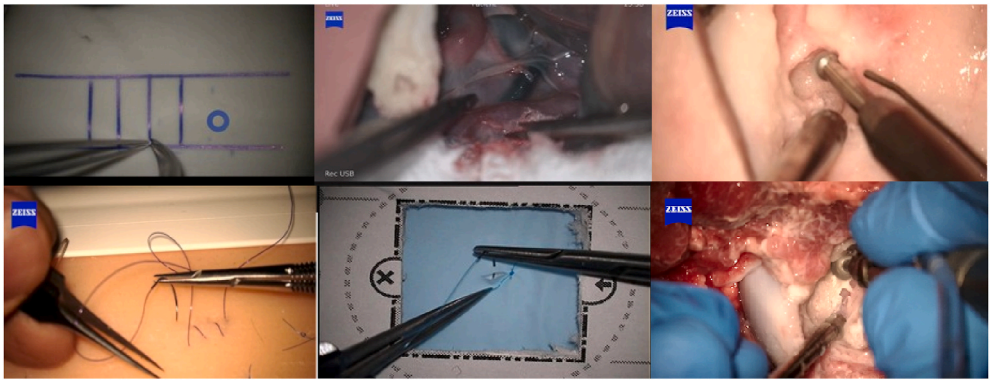

Fig. 1. Sample images from simulation training tasks used for training the network. Clockwise from top left: Interrupted suturing, vessel dissection, porcine drilling (outside), porcine drilling (inside), interrupted suturing, and continuous suturing.

single OR video may have included scenes from skull drilling and operations on the brain tissues. A total of 17 different instrument types were annotated with the goal of creating a large training dataset with diverse tools used in various scenarios (see Table 2). Thus, our dataset is more versatile compared to studies that merely focused on dissectors, scissors, and forceps in microsurgical procedures [11] or those with 14 different annotated objects from laparoscopic gastrectomy videos [59].

Frames were extracted at 0.5 to 15 frames per second (FPS) depending on the specifications of the source videos and recorded tasks. Consecutive frames with no visible movement were removed. The aim was to create a dataset where different tools were represented in as many settings as possible.

For training and validation, a total of 7500 frames were annotated using the LabelImg annotation tool [56]. Tools such as microforceps and microscissors were annotated with a bounding box that covers enough of the tool tip that it is possible to determine within the bounding box what the tool was doing — grasping or cutting something, for example (See Fig. 2). The needle and other smaller tools were annotated with a bounding box that contained the entire tool. For testing purposes, 1213 frames were annotated of which 2051 were obtained from the simulated procedures and the rest from the OR videos.

2.3. Description of experiments

Three experiments were conducted to train and evaluate the models. In the first two experiments, the model was trained and tested with data from videos of microsurgical training tasks. In the third experiment, data were added from real OR recordings to test the model's performance using real microsurgical scenarios.

For the first experiment, Exp. 1, two videos were set aside for the purpose of evaluating the model's performance in unseen settings. These videos showed a needle piercing and a porcine drilling task in three different settings that were similar, but not identical, to the videos used for model training and validation. In the needle piercing task, the unseen video showed a new training platform while in the porcine drilling task the drilling location was changed. The purpose was to test how well the detection works when small to moderate changes in the setting are introduced to the data.

For the second experiment, Exp. 2, we included frames from the two excluded videos into the training and validation datasets, and continued to train the model with the weights from the last epoch of Exp. 1. Although the new training frames were obtained from the same videos that were used for testing, they were not part of the original test frames. In Exp. 2, our goal was to investigate how the performance improves after adding a modest amount of new training data from the previously unseen settings. The time needed to annotate the new frames was less than 2 h.

For the third experiment, Exp. 3, we trained the model first with only frames from the OR videos and then again after adding frames from simulated microsurgery training sessions used in Exp. 1. The aim of this experiment was to test if the detection performance on OR surgical procedures could be improved by supplementing the training data consisting of OR recordings with a base dataset from simulation training environments. Using this base dataset has the potential to reduce the amount of annotated data required to detect tools in new settings.

2.4. Model training and evaluation

Table 2 shows the number of instances and frames used for the train, validation and test sets. The model was trained in Google Colab using Tesla V100-SXM2-16GB. In the first and third experiments the models were trained from the weights that had been pre-trained using the COCO dataset, provided in the YOLOv5 repository. The training was run for

Table 2
The number of annotated objects for each class in the training, validation, and test sets. Exp. 1 and 2 included frames from the simulation training tasks while frames from OR procedures were added for Exp. 3. In experiments 2 and 3, the number of training and validation instances and frames were added to the datasets used in Exp 1. For example, the Exp. 2 training dataset had 3101 + 55 instances of microforceps (see Table 1).

Experiment	Train			Validation			Test	
	1	2	3	1	2	3	1 & 2	3
Microforceps	3101	55	253	775	99	60	265	339
Needle	1799	22	24	439	48	6	118	77
Needleholder	2079	23	73	504	53	16	169	143
Microscissors	1379	33	225	362	45	68	100	194
Suction	1215	131	571	217	165	135	688	619
Drill	808	85	31	145	98	4	450	97
Hook	229	10	35	38	17	9	190	22
Kerrison	147	32	0	22	52	0	107	0
Knot/loop	2554	0	20	661	0	8	0	28
Scalpel			53			12		58
Water irrigation			17			3		30
Clamp applier			13			5		18
Clamp			204			60		93
Dissector			58			13		23
Bipolar			154			24		160
Aspirator			29			7		58
Retractor			153			34		94
Total	13,311	391	1913	3163	577	464	2087	2053
Frames	4931	231	748	1131	277	182	1213	838

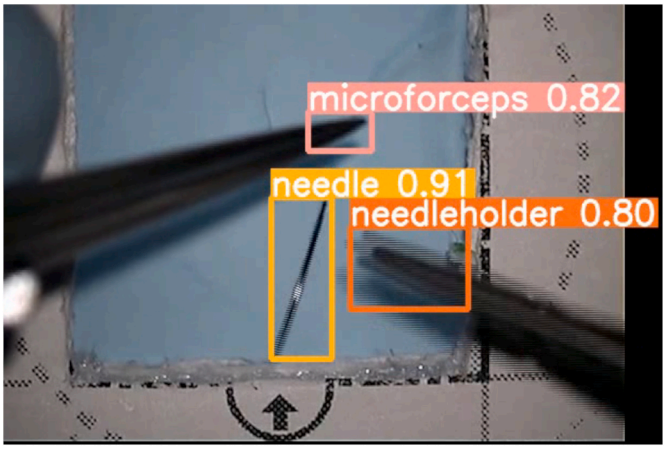

Fig. 2. An example of detection results from the interrupted suturing training task with three tools: microforceps, needle, and needleholder. Numbers in the bounding boxes indicate the detection confidence level.

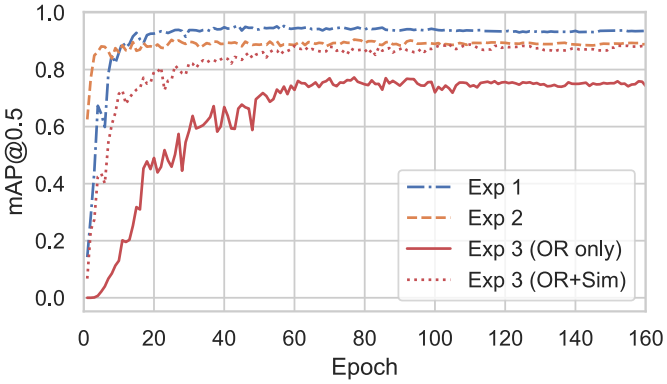

Fig. 3. Evolution of detection performance on the validation data during the first 160 training epochs. In Experiments 1 and 2, the network was trained only with the simulated training recordings. Results for Experiment 3 are presented for two cases; one using only the OR videos and the other when both OR and simulation tasks were used for model training and evaluation (see Table 1). Because the datasets differed in size, the epochs for different experiments involved different number of training iterations.

160 to 200 epochs with a minibatch size of 28 (approximately 28,000 to 42,000 iterations) and using the default hyperparameters provided in the YOLOv5 repository.[1] Once the training was complete, model weights that had resulted in the highest default fitness value over the validation set were saved. Fig. 3 demonstrates the mean average precision at the IoU threshold of 0.5 (mAP@0.5) on the validation sets of experiments 1 and 2 and the two cases of experiment 3.

The training reached a plateau after approximately 60–80 epochs for all three experiments.

Table 3 shows the validation and test mAPs for each tool class obtained at the IoU of 0.5. The last row denotes the overall mAPs for each experiment. The Kerrison rangeur was absent from the OR videos and was thus also from the test set of Exp. 3. Furthermore, Fig. 4 demonstrates the precision and recall-F1-score curves and break-even confidence thresholds for test sets of the three experiments. Generally speaking, all the test performance metrics improved when the model in Exp. 1 was continued to be trained for Exp. 2. For the validation datasets the F1 values at the break-even points were 0.930, 0.916 and 0.900 for Experiments 1–3, respectively.Results indicate that the average per-class mAP increased by 5.02% between Exp. 1 and 2, and the best detection precision is obtained for the drilling instruments. Furthermore, the test mAP improved for microforceps, microscissors, suction, and hook classes when new training videos were added in Exp. 2 and training continued from the last checkpoint. In fact, the hook observes the largest improvement of 27.44% in its mAP when only 10 new frames are utilized for retraining the model. This record is followed by the 10.45% improvement in the test set classification for the microscissors.

The last two columns of Table 3 demonstrate the classification results for Exp. 3 when the network was trained from scratch, first with only frames from the OR videos and again after adding frames from simulated microsurgery sessions. The test set mAPs are presented separately for these two cases. Besides kerrison that was not present in the test frames of Exp. 3, adding data from surgical training tasks improved or maintained the good detection results for 10 out of 16 remaining classes of

[1] https://github.com/ultralytics/yolov5/tree/master/data.

Table 3
Surgical tool classification results from Experiments 1–3 reported as per class average precision (AP) and overall mean AP at 0.5 IoU threshold (mAP@0.5). Test and validation results for Exp. 3 show the performance for models that were trained either only with the operation room (OR) data or with both the OR and simulation training task data (OR + Sim). The Exp. 3 validation results are given for frames from the OR videos.

Tool	Exp 1		Exp 2		Exp 3 (OR/OR + Sim)	
	Val.	Test	Val.	Test	Val.	Test
Microforceps	0.945	0.800	0.867	0.854	0.899/ 0.914	0.696/ 0.676
Needleholder	0.921	0.811	0.820	0.803	0.710/ 0.871	0.534/ 0.615
Needle	0.902	0.615	0.931	0.586	0/0	0.322/ 0.282
Microscissors	0.964	0.797	0.849	0.880	0.764/ 0.799	0.633/ 0.661
Suction	0.995	0.813	0.994	0.872	0.861/ 0.881	0.739/ 0.770
Drill	0.916	0.917	0.964	0.922	0.995/ 0.995	0.995/ 0.995
Hook	0.834	0.441	0.775	0.562	0.720/ 0.714	0.887/ 0.914
Kerrison	0.843	0.384	0.927	0.374	−/−	−/−
Knot/loop	0.934	–	0.930	–	−/−	0.340/ 0.363
Scalpel					0.687/ 0.765	0.653/ 0.725
Water irrigation					0.686/ 0.673	0.247/ 0.312
Clamp applier					0.920/ 0.995	0.378/ 0.313
Clamp					0.835/ 0.802	0.615/ 0.555
Dissector					0.734/ 0.680	0.467/ 0.378
Bipolar forceps					0.981/ 0.979	0.672/ 0.741
Aspirator					0.995/ 0.995	0.913/ 0.873
Retractor					0.722/ 0.737	0.634/ 0.662
mAP@0.5	**0.914**	**0.697**	**0.895**	**0.732**	**0.767/ 0.788**	**0.608/ 0.615**

microsurgical tools, namely the needleholder, microscissors, suction, drill, hook, knot and loop, scalpel, water irrigation, bipolar forceps, and retractors. Furthermore, the overall test detection improved by 1.15% between these two cases. The Supplementary Material includes a compilation of three video clips from the real OR surgeries that demonstrate successful detections for Exp. 3.

The best detection results in Table 3 are obtained for the most frequent tools with the largest amount of training data, namely the microforceps, microscissors, needleholder, and the suction device (Table 2); needle detection results are notably poor, especially in the real surgical videos, despite having a similar amount of training data as microscissors and needle holders. One explanation is that the needle is often occluded and can appear in many different poses, thus requiring even more training data for reliable detection. Furthermore, the OR videos used in experiment 3 were heavily edited at the source to focus on the main task in the procedure, which rarely required the needle and needleholder, thus further limiting the amount of data available for training and testing in our experiments.

2.4.1. Inference time

When run within Google Colab using Tesla P100-PCIE-16GB, the average inference time for a single frame with the resolution of 1280 × 720 pixels was 25.04 ± 0.75 ms or 39.93 frames per second on the test sets. When the detection was run on mp4-format videos with the resolution of 1280 × 720 pixels the inference time was 25.07 ± 0.69 ms per frame, corresponding to 39.89 frames per second. This extremely short inference time, which is comparable with the mean detection time of 23 ms reported by Zhao et al. [61], enables the model to detect surgical instruments in real time with minimal latency.

3. Case study: eye-hand coordination in microsurgical vessel dissection

Next we describe a case study in which the tool detection model is combined with an eye tracker in a novel microsurgical application. We analyze the associations of task ergonomics on tool detection and investigate the system's suitability for extracting tool kinematics from microsurgical videos. We demonstrate that, by analyzing concurrent recordings of tool and gaze movements, it is possible to distinguish behavioral differences between surgical actions.

3.1. Study setup

An experienced neurosurgeon conducted a simulated intracranial vessel dissection using a surgical microscope (Carl Zeiss Meditec) that was equipped with an add-on eye tracker. One camera recorded the operating field through the microscope while a ceiling-mounted camera recorded the other activities in the room. The task was performed in two

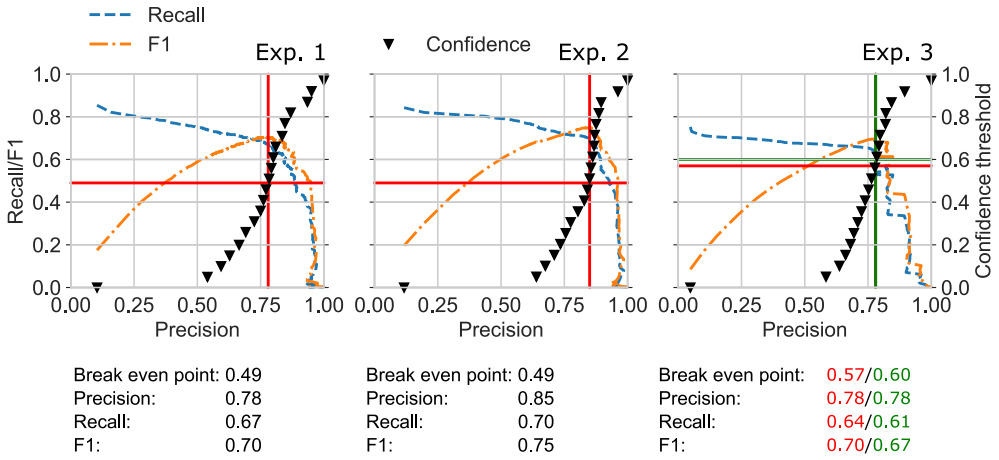

Fig. 4. Precision, recall, and F1-score at different confidence thresholds for the test sets in Experiments 1–3. Exp. 3 results are given for the model trained only with OR data (in red) and the model trained with OR and simulation data (in green). The break-even point indicates the confidence threshold (y-axis, right) with the highest F1 value (y-axis, left), and is indicated in the figure by the horizontal and vertical lines. The listed precision, recall and F1 values are for this confidence threshold.

sessions, referred to as Session 1 and Session 2, and both sessions lasted about 20 min. Session 1 was performed with a small cranial opening that represented a minimally invasive surgical opening, whereas Session 2 was performed with a wider opening that represented a classical cranial opening. These different setups motivate the surgeon to employ different surgical and ergonomical strategies, as they require varying levels of dexterity and eye-hand coordination. In both sessions, the neurosurgeon used microforceps and microscissors to complete the task.

3.1.1. Eye-tracker setup

A number of practical solutions exist for eye-tracking in microsurgery [12,14]. We utilized a See True Technologies[2] add-on eye tracker, version 2020. This eye tracker is based on video-oculography and is attached securely on the top of oculars without reducing the surgeon's field of view. With a sampling rate of 30 Hz, the eye tracker provides pupil detection and gaze-to-scene mapping in real time and stores the data in videos and timestamped files. At the beginning of each dissection session, the eye tracker was calibrated with a 9-point target placed on top of the opening through which the dissection was performed.

3.1.2. Microsurgical action annotation

Microscope videos were segmented into four microsurgical actions: Dissection (D), where the vessel was dissected using the microscissors; enhancement of the visual scene (EVS), where the tools were used to move the tissue or other objects to reveal the dissection area; exploration with the tools (ET), where the microscissors or microforceps were used to find a new dissection location and explore the relative anatomy; and intervention (I), where a surgical gauze was used to clean unexpected bleeding. The first and last frames of each action were annotated using the Behavioral Observation Research Interactive Software (BORIS) [16]. Fig. 5 demonstrates example frames for each of the four actions.

Next, continuous work phases were annotated by removing segments without surgical actions. Such segments occurred when the surgeon adjusted the microscope or switched surgical instruments. Within each continuous work phase, the four aforementioned actions could repeat several times.

3.1.3. Kinematics computation and pipeline

The tool detection model described in Section 2 was applied to the microscope scene recordings. From the detected tool tip coordinates, we computed six kinematic metrics: path length (PL), velocity, acceleration, jerk, curvature, and the inter-tool tip distance (TD), inspired by Davids et al. [11]. Path length was calculated as the sum of frame-to-frame displacements of the tool tips within each action. Velocity was measured as the frame-to-frame displacement of the, tool tips. Acceleration and jerk were calculated as absolute values of the first and second derivatives of the instantaneous velocity, i.e. the displacement of the tool tips between two frames. Curvature was calculated by

$$c = \frac{|\overline{a}_i \times \overline{v}_i|}{v_i^3},$$

where $\overline{a}$ and $\overline{v}$ are the acceleration and velocity vectors, respectively, and v_i is the magnitude of the instantaneous velocity in frame i.

Inter-tool tip distance was calculated as the Euclidean distance between the tool tips when both tools were present in the same frame. These metrics were extracted separately for each continuous work phase as explained in Section 3.1.2.

Fig. 6 shows a pipeline developed for processing the tool kinematic data for each continuous work phase. Steps (1) and (2) dealt with the missed detection of surgical instruments. Outliers were removed in step (3): the histogram of the tools' instantaneous velocity were plotted for the entire recordings, and outliers were compared with the original videos. It was determined that changes in tool tip coordinates that surpassed 50 pixels were either noise or caused by faulty detection. Kinematic metrics were calculated from continuous actions in step (4). To avoid reporting biased measurements caused by changes in microscope magnification levels, mean tool metrics were normalized by dividing them with the largest calculated mean value for each magnification level and tool as shown in steps (5) and (6). Addressing variations in magnification ratios and microscope movements are a novelty of the current paper and was not included in the previous studies on tool kinematic analysis in microsurgery [11].

3.1.4. Frame-based gaze-tool distance analysis

Distances between gaze and the tools or the movement targets have been shown to depend on the subject's prior experience with the task [4,28,31]. Therefore, obtaining accurate gaze-tool distance measures is central to evaluating surgical proficiency and motor control strategies. To investigate if the gaze-tool distance differs between actions, we calculated the pixel-wise distances between the gaze coordinates and centers of the bounding boxes for frames in which the gaze and at least one instrument were detected (Ref. Sections 3.1.1 and 3.1.3). The eye tracker data was processed to remove faulty detections, such as when the gaze was mapped outside the frame. If fewer than five consecutive frames were missing, their data was computed with linear interpolation. Actions with more than 20% of missing data were removed.

3.1.5. Statistical analysis

Differences in gaze-tool distances among actions were analyzed as follows. The normalized mean and standard deviation (SD) of tool velocity, acceleration, jerk, curvature, and inter-tool tip distance were computed for each continuous microsurgical action. For each session and tool, an independent, two-sided Wilcoxon rank sum test (Mann-Whitney U test) for equal medians was applied to study the effect of microsurgical actions on the continuous actions. Furthermore, tool metrics from all continuous actions within each session were grouped, and their pairwise Pearson's linear correlation as well as critical values for correlation coefficients were computed. Statistical analysis was performed using MATLAB, version R2020b (MathWorks Inc., Natick, Massachusetts, USA), and R, version 3.6.0 [47].

3.2. Results

Annotating the microsurgical scene videos resulted in a total of 71 actions for Session 1 with a small cranial opening and 96 actions annotated for Session 2 with a wider opening. The total duration of actions in Session 1 was 25,474 frames, from which 20,363 frames (79.94%) had at least one successful tool detection. Session 2 had a total duration of 25,835 frames where 23,792 frames (92.09%) had at least one detection. Once the processing pipeline of Fig. 6 was applied on the annotated microsurgical actions, 27 and 24 microsurgical actions were removed from Session 1 and Session 2, respectively. After step 3, 55 frames were removed from a total of 22,115 frames in Session 1 for which at least one instrument was detected. Likewise, 186 frames were removed from 25,097 frames in Session 2. The final number of continuous actions was 44 (D: 27, EVS: 15, ET: 2, I: 0) for Session 1 and 74 (D: 37, EVS: 28, ET: 3, I: 6) for Session 2. The Supplementary Material includes a video clip from Session 2 demonstrating the precise detection of gaze, microforceps, and microscissors.

The median duration of the remaining actions was 7.4 ± 11.7 s for Session 1 and 6.1 ± 9.1 s for Session 2. Table 4 shows the distribution of frames from both sessions that included gaze and at least one detected instrument. It can be seen that dissection and enhancement of the visual scene accounted for around 70% and 20% of the total task duration, respectively.

3.2.1. Normalized tool kinematics in intracranial vessel dissection

Table 5 presents the medians of each metric for D and EVS actions

[2] https://www.seetruetechnologies.com/.

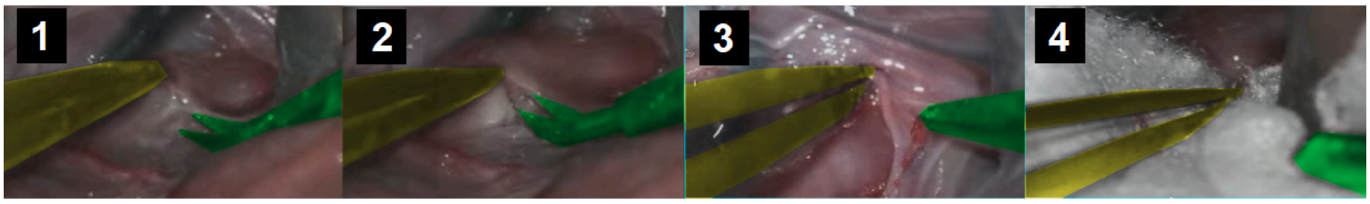

Fig. 5. Microsurgical actions annotated from microscope scene videos. (1) EVS: Microforceps, shown in yellow, are used to lift the tissue to clear the path in the dissection area. (2) D: Microscissors, shown in green, are used to dissect the tissue. (3) ET: Microforceps are used to lift the outer layer of the tissue to find a dissection location. (4) I: A surgical gauze is applied to clean blood.

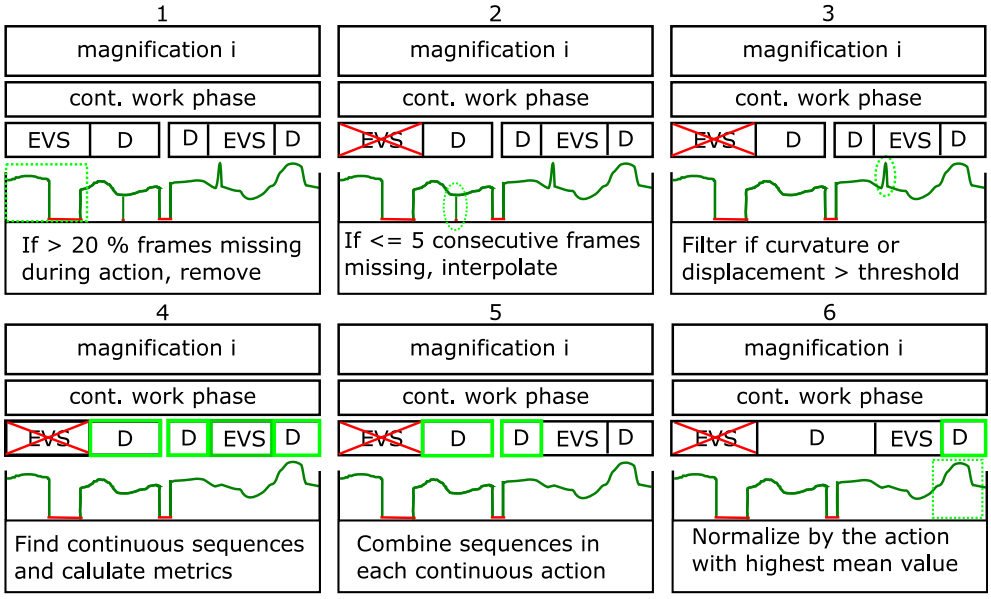

Fig. 6. Tool kinematic data processing pipeline. Under each magnification level, there were one or more continuous work phases, which consisted of several continuous actions. The green curve represents the tool's velocity, and the red lines are missing frames.

Table 4
Number of frames with detected gaze and instrument across four different microsurgical actions. D: dissection; EVS: enhancement of the visual scene; ET: exploration with the tools; I: intervention.

Instrument	Action	# Detected Frames (% of total)	
		Microforceps	Microscissors
Session 1	EVS	2740 (16.1)	2577 (15.4)
	D	12,832 (75.6)	12,859 (76.8)
	ET	1400 (8.2)	1302 (7.8)
	I	–	–
Session 2	EVS	4944 (23.0)	4908 (22.2)
	D	13,410 (62.5)	13,576 (61.5)
	ET	1655 (7.7)	2190 (10.0)
	I	1441 (6.7)	1412 (6.4)

and the p-values for the two-sided Wilcoxon rank sum tests for Sessions 1 and 2. Table 6 includes the results for mean and SD values of inter-tool tip distances. Due to the small number of intervention and ET actions (Ref. Table 4), only dissection and EVS were included in the analysis. The family-wise error rate (FWER) was computed at the 0.05 significance level for 40 conducted hypothesis tests. Results indicate that, in both sessions, the mean and SD of velocity, acceleration, and jerk of microforceps were significantly larger during the EVS than tissue dissection, $p < 0.001$ 2. For microscissors, only the normalized path length in Session 1 was significantly larger during dissection, $p < 0.001$ 2. Furthermore, the normalized mean and SD of inter-tool tip distance were significantly larger during the EVS, $p < 0.01$, in Session 2 that allowed more freedom of movement to the surgeon.

Finally, Fig. 7 displays heatmaps of significant tool kinematic correlations. The normalized mean and SD of velocity, acceleration, and jerk show large positive correlations for both microforceps and microscissors. However, the small but positive correlations of curvature and tool distance SD with other metrics are only observed for microforceps in Session 1. In Session 2, all metrics except for normalized mean curvature demonstrate significant positive correlations with other metrics.

3.2.2 Tool. Gaze- distance analysis in intracranial vessel dissection

Results of two-sided, independent Mann-Whitney U tests to compare gaze-tool distances between dissection and EVS actions are reported in Table 7. The FWER was computed at the 0.05 significance level for 8 conducted hypothesis tests. As shown in Table 7, none of the gaze-tool metrics were significantly different between D and EVS at the FWER significance level of 0.006. However, the normalized gaze-microscissors distances were smaller during dissection than EVS, $p < 0.01$, in Session 2 that had provided the neurosurgeon with a larger space to maneuver.

4. Discussion

In this work, we trained and evaluated a fast, CNN-based detector to perform tool detection from a dataset with approximately 20 surgical settings and 17 microinstruments. To demonstrate the feasibility of using this model analysing microsurgical performance, we introduced a tool kinematic processing pipeline and fused the tool detections with gaze tracking to monitor tool kinematics and eye-hand coordination during a highly magnified microsurgical task.

Table 5
Median values of normalized tool kinematics and results of two-sided Mann-Whitney U tests for Session 1 and 2. Asterisks denote $p < 0.001$ 2. D: dissection; EVS: enhancement of the visual scene; SD: standard deviation.

Session 1

	Microforceps			Microscissors		
Tool motion metrics	D	EVS	p	D	EVS	p
Norm. path length	0.282	0.147	0.121 4	0.257	0.073	0.001 1*
Norm. mean velocity	0.385	0.864	0.000 2*	0.590	0.575	0.545 9
Norm. SD velocity	0.344	0.905	0.000 1*	0.571	0.591	0.916 4
Norm. mean acceleration	0.303	0.706	0.000 0*	0.543	0.579	0.772 7
Norm. SD acceleration	0.407	1.030	0.000 2*	0.694	0.604	0.833 7
Norm. mean jerk	0.306	0.819	0.000 2*	0.569	0.588	0.823 4
Norm. SD jerk	0.424	1.084	0.000 3*	0.713	0.590	0.618 0
Norm. mean curvature	0.727	0.665	0.109 3	0.654	0.765	0.031 3
Norm. SD curvature	1.747	1.762	0.693 8	1.883	2.036	0.318 5

Session 2

	Microforceps			Microscissors		
Tool motion metrics	D	EVS	p	D	EVS	p
Norm. path length	0.178	0.284	0.705 7	0.202	0.137	0.143 2
Norm. mean velocity	0.254	0.453	0.000 0*	0.328	0.373	0.201 1
Norm. SD velocity	0.300	0.510	0.000 0*	0.334	0.431	0.093 8
Norm. mean acceleration	0.238	0.440	0.000 0*	0.331	0.410	0.150 6
Norm. SD acceleration	0.323	0.554	0.000 3*	0.423	0.473	0.158 3
Norm. mean jerk	0.258	0.457	0.000 0*	0.283	0.340	0.129 3
Norm. SD jerk	0.328	0.578	0.000 4*	0.386	0.434	0.170 3
Norm. mean curvature	0.775	0.652	0.002 7	0.608	0.594	0.878 9
Norm. SD curvature	1.673	1.697	0.755 6	1.702	2.011	0.150 6

Table 6
Median values for normalized inter-tool tip distances and results of two-sided Mann-Whitney U tests for sessions 1 and 2. Asterisk denotes $p < 0.001$ 2. MF: microforceps; MS: microscissors.

Tool motion metrics	Session 1: MF-MS			Session 2: MF-MS		
	D	EVS	p	D	EVS	p
Norm. mean tool tip distance	0.607	0.580	0.511 6	0.546	0.634	0.002 8
Norm. SD tool tip distance	0.114	0.113	0.979 1	0.095	0.152	0.000 3*

4.1. Tool detection validity

As shown in Table 3, the microsurgical tool detection performed well on the validation datasets in each of the three tested experiments. These results show that YOLOv5-l is well suited for tool detection from simulated training tasks and for datasets with several recordings from microsurgical procedures. Our validation results for mAP@0.5 were 89.5% for Experiment 2 and 91.4% for Experiment 1. These results are highly comparable with those of Yamazaki et al. [59], who achieved an average validation precision of 91.81% and test precision of 83.75% with the YOLOv3 detector on a dataset of laparoscopic gastronomies.

Cho et al. used YOLOv2 and RetinaNet for surgical tool tip detection in an endoscopic spine surgery task, and reported F1-scores of 74.7%–84.6% for RetinaNet and 76.6%–83.5% for YOLOv2 [7]. However, their dataset was highly homogeneous with only one instrument recorded from one type of procedure and their aim was detecting the very tip of the tool. As shown in Fig. 4 and Section 2.4, the YOLOv5-l model achieved F1-scores as high as 93.0% on the validation and 75.0% on the test sets composed of simulated and real microsurgical procedures with multiple instruments.

4.2. Performance in unseen simulation data

Results for the test dataset of Exp. 1 in Table 3 demonstrate that the performance remained high for more common tools, particularly in the piercing task that was performed using microscissors, microforceps, and a needle against a plain background. However, differences between validation and test results also show that the detection can be sensitive to changes in the surgical settings, even when the changes are small. Comparing the number of annotated instruments in Table 2 and detection performance in Table 3 indicates that this reduction was largest for the tools that were used infrequently and thus had less available training data.

The results indicate that tool detection models can be trained to perform well in a particular surgical procedure, as also reported in previous research, but still exhibit performance degradation when changes are introduced. Developing even more robust models for surgical tool detection requires collecting more diverse datasets of tools and surgical procedures. Our contribution stimulates such efforts and we invite the community to compile multi-site, multi-procedure datasets of microsurgical images with and without tools.

4.3. Performance in real microsurgical procedures

As mentioned in Section 1.3, Exp. 3 was conducted to investigate transferability of learning between simulated and real microsurgical scenarios. Results of Exp. 3 (Table 3) concerning the more common tools (microforceps, microscissors, and suction) show that the model can achieve very good performance in real OR videos even with limited amount of data. The performance remained sensitive to changes because of the small size of the dataset; however, extending the dataset with recordings from simulated training tasks further improved tool detection in real setups in 10 out of 16 tools. Furthermore, the overall performance in tool detection of OR videos improved by incorporating the knowledge learned from simulation training tasks, demonstrating the first successful implementation of transfer learning in surgical tool detection.

4.4. Inference time

As mentioned in Section 2.4.1, our model obtained the average inference time of 25.04 ± 0.75 and 25.07 ± 0.69 ms/frame when tested on single frames and mp4 videos, respectively, corresponding to detecting 39.90 ± 1.2 frames per second. These results are comparable with the 23-ms detection time of [61] for ATLAS Dione and Endovis Challenge datasets, and outperform the best detection rates of 26.51 frames/s from the CNN-based model in Ref. [62] and 25 frames/s of YOLOv3 in Ref. [59].

4.5. Interpreting the tool kinematics

To demonstrate the feasibility of the tool detection model in microsurgical training tasks, we applied the gaze and tool detection pipelines in an intracranial vessel dissection case study. Few prior studies have focused on microsurgery, and to the best of our knowledge, Chainey et al. [6] is the only work that has examined eye-hand coordination in microsurgery. However, they evaluated performance with temporal metrics, such as the delay between gaze and tool movements, and not directly with tool kinematics.

Statistically significant differences in microforceps kinematics between dissection and EVS demonstrate the possibility of distinguishing the surgeon's actions (Section 3.2.1). Using similar metrics, extracted with a mask region CNN, Davids et al. were able to separate novice and expert surgeons in a simulated arachnoid dissection task [11]. They found significant differences in means and standard deviations of

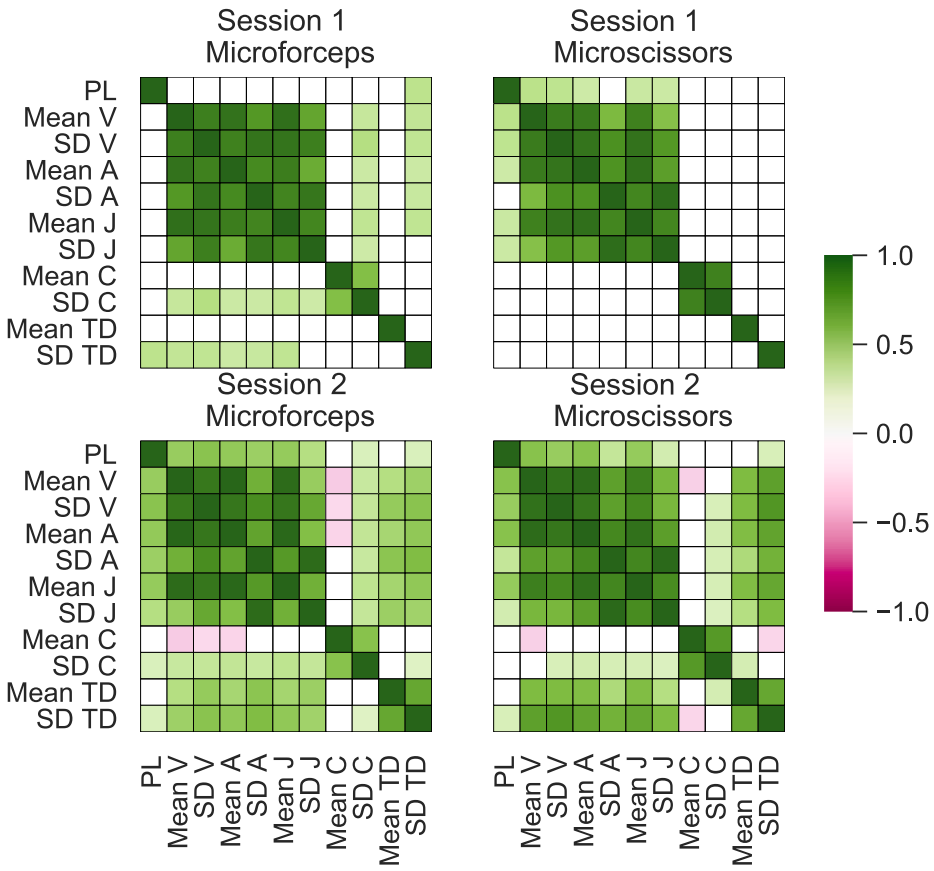

Fig. 7. Correlation heatmaps for normalized tool kinematics computed from continuous surgical actions. Correlations smaller than the critical values at 0.05 significance level have been masked to 0. Tool kinematics are indicated with the abbreviations PL: path length, V: velocity, A: acceleration, J: jerk, C: curvature, and TD: tool-tip distance.

Table 7
Median values for normalized gaze-tool metrics of individual tools and actions and results of Mann-Whitney U tests for sessions 1 and 2. MF: Microforceps; MS: Microscissors.

	Session 1: MF			Session 1: MS		
	D	EVS	p	D	EVS	p
Gaze-Tool Mean	0.516	0.573	0.512 3	0.756	0.834	0.341 5
Gaze-Tool SD	0.147	0.160	0.568 4	0.186	0.140	0.363 6
	Session 2: MF			Session 2: MS		
	D	EVS	p	D	EVS	p
Gaze-Tool Mean	0.566	0.480	0.187 3	0.536	0.714	0.008 6
Gaze-Tool SD	0.126	0.182	0.089 7	0.151	0.196	0.187 4

velocity, jerk, curvature, and the standard deviation of acceleration. In our results (Table 5), the same metrics – with the exception of the curvature – revealed statistically significant differences between microsurgical actions. Notably, the differences were significant for the microforceps, the left hand tool, but not for the microscissors. The result is attributable to the nature of the dissection task, since microscissors move less even when actively used, whereas microforceps move faster and at higher accelerations when enhancing the visual scene.

Davids et al. introduced the inter-tool tip distance to evaluate surgical skills [11]. This metric could distinguish between actions with the larger opening in Session 2, but not with the smaller opening in Session 1 (Table 6). A possible explanation is that the smaller opening affords less space for tool movements and thus limits the possible differences in inter-tool tip distances.

Our analysis also demonstrated strong linear correlations among the means of velocity, acceleration, jerk, and their standard deviations (Fig. 7) that were not reported in other studies. These metrics are partially correlated by definition: for example, performing an action with a high jerk increases the likelihood of higher acceleration. Interestingly, the correlation between inter-tool tip distance and other metrics was stronger in Session 2 than in Session 1, likely because the former provided more space for independent and between-tool movements.

4.6. Monitoring gaze-tool distance and eye-hand coordination

Our results demonstrate that only the mean gaze-microscissors distance was significantly different between dissection and EVS in Session 2. For other tool-session combinations, the gaze-tool distance did not change significantly (Table 7). Therefore, the mean gaze-tool distance may not be a suitable metric for separating these particular actions and segmentation granularity. Previously, a fusion of eye and hand metrics was used to classify longer surgical phases in laparoscopy [53]. The applicability of similar feature selection and classification approaches should be explored in microsurgery.

4.7. Challenges and future directions

Fig. 8 demonstrates six examples of challenging tool detection situations. Besides misclassifications, as in plot A, tools can be temporarily occluded by the surgeons' hands, other instruments, and tissues (plots B, C, and D). Images can also be out of focus because of high magnification as in plot E. Partial or complete occlusion is more common for specific tools such as the needles. These issues are worsened during actual surgical procedures where the tools are often manipulated in extremely confined spaces.

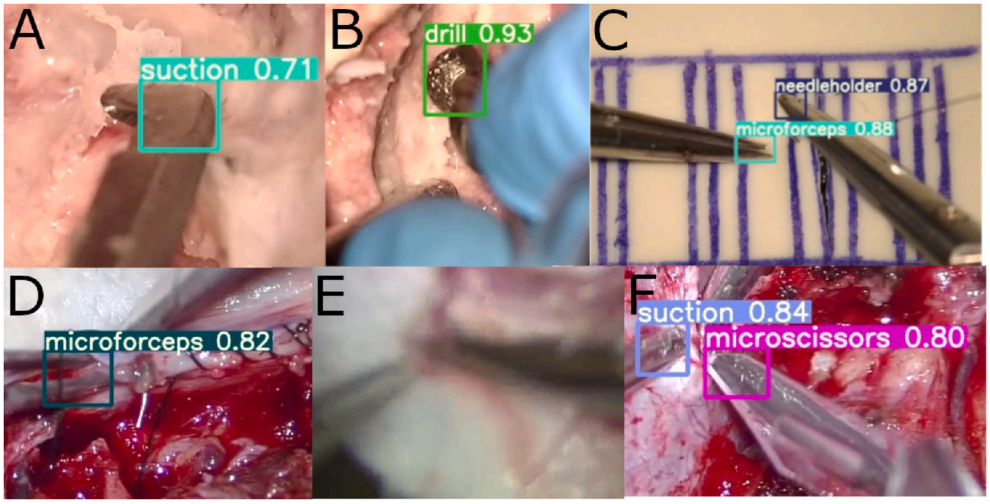

Fig. 8. Examples of challenging situations. *Top row*: frames from surgical training tasks. (A) Kerrison rangeur misclassified as the suction device. (B) Suction device not detected due to occlusion by the glove. (C) Needle not detected due to occlusion by the needleholder and the latex surface. *Bottom row*: frames from actual surgical operations. (D) Needleholder and needle not detected. (E) Microforceps and needleholder not detected due to extremely low focus. (F) Scalpel misclassified as microscissors.

Furthermore, the surgical approach may restrict the visibility to the scene, such as in Session 1 of our case study where a narrow opening was used to perform the task. The narrow opening leads to less illumination getting to the scene and more movements that are coaxial with the camera and thus cannot be tracked as accurately. The coaxial movements are also more likely to result in blurred tool tips.

The tool detection models were trained with 4900 to 5900 frames from about 20 different settings. The best performance was seen with instruments that had at least 1000 training samples (see Table 2). As different anatomies, procedures, and instruments introduce new sources of variability, a computer vision-based detector that performs robustly under all circumstances requires more training data from diverse surgical settings.

Finally, accurate calculation of tool movement metrics requires converting the pixel values into physical units. Except for surgical training tasks in which the microscope magnification and pose can be fixed, performing such conversion is problematic. Previous work did not address the magnification ratio and distance calculations under different optical zoom ratios [11]. We resolved this problem by normalizing the calculated metrics by the largest mean value under each magnification level. General automated methods for correcting microscope movements should be explored in the future.

5. Conclusions

A new deep learning model was presented and evaluated for automatic microsurgical tool detection in simulation training tasks and intraoperative videos. The feasibility of monitoring tool kinematics and eye-hand coordination across different microsurgical actions was demonstrated in a case study of a simulated intracranial vessel dissection task.

The presented methodology for eye-hand coordination analysis eliminates the need for additional optical and magnetic tracking sensors that suffer from inherent interference and positional drifts [46] as well as the need for manual annotations. This pipeline will assist educators and researchers with automated content extraction [46] and objective assessment of bimanual tool handling during skill acquisition in microsurgery [1,22,52].

Funding

The research was supported by the Academy of Finland grant #34658.

Data statement

The tool detection is based on an open source model, available at https://github.com/ultralytics/yolov5. The code written to process the case study data is available at https://github.com/jpkos/GazeTooIAnalysis. Data used to train the detection model and data from the case study are partially available from the authors upon request.

Declaration of competing interest

Prof. Bednarik declares affiliation with an eye-tracking company, however no financial or legal interests that would affect the current study. Other authors declare no conflicts of interest.

Appendix A. Supplementary data

The Supplementary Material contains a compilation of three videos that were the result of model predictions in Experiment 3. A second clip is provided to demonstrate the precise detection of gaze, microforceps, and microscissors by the case study pipeline applied to Session 2 of the simulated intracrianial operation.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.105121.

References

[1] N. Ahmidi, G.D. Hager, L. Ishii, G. Fichtinger, G.L. Gallia, M. Ishii, Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2010, pp. 295–302, https://doi.org/10.1007/978-3-642-15711-0_37.

[2] R. Bednarik, P. Bartczak, H. Vrzakova, J. Koskinen, A.-P. Elomaa, A. Huotarinen, D. G. de Gómez Pérez, M. von und zu Fraunberg, Pupil size as an indicator of visual-motor workload and expertise in microsurgical training tasks, in: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, 2018, pp. 1–5, https://doi.org/10.1145/3204493.3204577.

[3] E. Belykh, N.R. Onaka, I.T. Abramov, K. Yağmurlu, V.A. Byvaltsev, R.F. Spetzler, P. Nakaj, M.C. Preul, Systematic review of factors influencing surgical performance: practical recommendations for microsurgical procedures in neurosurgery, World Neurosurg. 112 (2018) e182–e207, https://doi.org/10.1016/j.wneu.2018.01.005.

[4] T.J. Bosch, T. Hanna, K.A. Fercho, L.A. Baugh, Behavioral performance and visual strategies during skill acquisition using a novel tool use motor learning task, Sci. Rep. 8 (2018) 1–11, https://doi.org/10.1038/s41598-018-32001-4.

[5] A.J. de Brouwer, J.R. Flanagan, M. Spering, Functional use of eye movements for an acting system, Trends Cognit. Sci. (2021) 252–263, https://doi.org/10.1016/j.tics.2020.12.006.

[6] J. Chainey, Examining Surgeon's Eye-Hand Coordination during Microsurgery, Master's thesis, University of Alberta, Edmonton, Canada, 2020, https://era.library.ualberta.ca/items/bcafa72a-408a-4d83-b3cc-080c4d38a5c1. (Accessed 14 September 2021).

[7] S.M. Cho, Y.G. Kim, J. Jeong, I. Kim, H.j. Lee, N. Kim, Automatic tip detection of surgical instruments in biportal endoscopic spine surgery, Comput. Biol. Med. 133 (2021) 104384, https://doi.org/10.1016/j.compbiomed.2021.104384.

[8] B. Choi, K. Jo, S. Choi, J. Choi, Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Ieee, 2017, pp. 1756–1759, https://doi.org/10.1109/EMBC.2017.8037183.

[9] J. Choque-Velasquez, D.A. Kozyrev, R. Colasanti, P. Thiarawat, P. Intarakhao, B. R. Jahromi, J. Hernesniemi, The open access video collection project "Hernesniemi's 1001 and more microsurgical videos of neurosurgery": a legacy for educational purposes, Surg. Neurol. Int. 8 (188) (2017), https://doi.org/10.4103/sni.sni_158_17.

[10] J.D. Crawford, W.P. Medendorp, J.J. Marotta, Spatial transformations for eye–hand coordination, J. Neurophysiol. 92 (1) (2004) 10–19, https://doi.org/10.1152/jn.00117.2004.

[11] J. Davids, S.G. Makariou, H. Ashrafian, A. Darzi, H.J. Marcus, S. Giannarou, Automated vision-based microsurgical skill analysis in neurosurgery using deep learning: development and preclinical validation, World Neurosurg. (2021) 669–686, https://doi.org/10.1016/j.wneu.2021.01.117.

[12] S. Eivazi, R. Bednarik, V. Leinonen, M. von und zu Fraunberg, J.E. Jääskeläinen, Embedding an eye tracker into a surgical microscope: requirements, design, and implementation, IEEE Sensor. J. 16 (2015) 2070–2078, https://doi.org/10.1109/JSEN.2015.2501237.

[13] S. Eivazi, A. Hafez, W. Fuhl, H. Afkari, E. Kasneci, M. Lehecka, R. Bednarik, Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope, Acta Neurochir. 159 (2017) 959–966, https://doi.org/10.1007/s00701-017-3185-1.

[14] Eivazi, S., Maurer, M., . Eyemic: an eye tracker for surgical microscope, in: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. doi:10.1145/3204493.3208342.

[15] J. Fooken, M. Spering, Eye movements as a readout of sensorimotor decision processes, J. Neurophysiol. 123 (2020) 1439–1447, https://doi.org/10.1152/jn.00622.2019.

[16] O. Friard, M. Gamba, Boris: a free, versatile open-source event-logging software for video/audio coding and live observations, Methods Ecol. Evol. 7 (2016) 1325–1330, https://doi.org/10.1111/2041-210X.12584.

[17] A. Ghanem, M. Kearns, A. Ballestín, S. Froschauer, Y. Akelina, S. Shurey, J. Legagneux, S. Ramachandran, S. Cozzolino, V. Ramakrishnan, G. Pafitanis, Y. Zakaria, K. Al-Maaytah, S. Komatsu, Y. Kimata, I. Cifuentes, P.N. Soucacos, P. Tos, S. Myers, International microsurgery simulation society (IMSS) consensus statement on the minimum standards for a basic microsurgery course, requirements for a microsurgical anastomosis global rating scale and minimum thresholds for training, Injury 51 (2020) S126–S130, https://doi.org/10.1016/j.injury.2020.02.004.

[18] A.C.P. Guédon, S.E.P. Meij, K.N.M.M.H. Osman, H.A. Kloosterman, K.J. van Stralen, M.C.M. Grimbergen, Q.A.J. Eijsbouts, J.J. van den Dobbelsteen, A. P. Twinanda, Deep learning for surgical phase recognition using endoscopic videos, Surg. Endosc. (2020) 6150–6157, https://doi.org/10.1007/s00464-020-08110-5.

[19] A. Harvey, J.N. Vickers, R. Snelgrove, M.F. Scott, S. Morrison, Expert surgeon's quiet eye and slowing down: expertise differences in performance and quiet eye duration during identification and dissection of the recurrent laryngeal nerve, Am. J. Surg. 207 (2014) 187–193, https://doi.org/10.1016/j.amjsurg.2013.07.033.

[20] M.M. Hayhoe, Vision and action, Annu. Rev. Vis. Sci. 3 (2017) 389–413, https://doi.org/10.1146/annurev-vision-102016-061437.

[21] D.S. Hersh, K.N. Sanford, K. Moore, F.A. Boop, Midline suboccipital craniotomy and direct stimulation for a dorsally exophytic brainstem tumor, Neurosurg. Focus: Video FOCVID 1 (2019) V9, https://doi.org/10.3171/2019.10.FocusVid.19456. (Accessed 12 March 2021).

[22] E.F. Hofstad, C. Våpenstad, M.K. Chmarra, T. Langø, E. Kuhry, R. Mårvik, A study of psychomotor skills in minimally invasive surgery: what differentiates expert and nonexpert performance, Surg. Endosc. Other Intervent. Tech. 27 (2013) 854–863, https://doi.org/10.1007/s00464-012-2524-9.

[23] Z. Huang, J. Wang, X. Fu, T. Yu, Y. Guo, R. Wang, Dc-spp-yolo: dense connection and spatial pyramid pooling based yolo for object detection, Inf. Sci. 522 (2020) 241–258, https://doi.org/10.1016/j.ins.2020.02.067.

[24] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, L. Fei-Fei, Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks, in: Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018 2018-Janua, 2018, pp. 691–699, https://doi.org/10.1109/WACV.2018.00081.

[25] K. Jo, Y. Choi, J. Choi, J.W. Chung, Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction, Appl. Sci. 9 (2019) 2865, https://doi.org/10.3390/app9142865.

[26] Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, yxNONG, Hogan, A., lorenzomammana, AlexWang1900, Chaurasia, A., Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Durgesh, Ingham, F., Frederik, Guilhen, Colmagro, A., Ye, H., Jacobsolawetz, Poznanski, J., Fang, J., Kim, J., Doan, K., Yu, L., 2021. Ultralytics/yolov5: v4.0 - nn.SiLU( ) activations, Weights & Biases logging, PyTorch Hub Integrat.. doi:10.5281/zenodo.4418161.

[27] R.S. Khan, G. Tien, M.S. Atkins, B. Zheng, O.N. Panton, A.T. Meneghetti, Analysis of eye gaze: do novice surgeons look at the same location as expert surgeons during a laparoscopic operation? Surg. Endosc. 26 (2012) 3536–3540, https://doi.org/10.1007/s00464-012-2400-7.

[28] J. Koskinen, R. Bednarik, Gaze-grabber distance in expert and novice forest machine operators: the effects of automatic boom control, in: ACM Symposium on Eye Tracking Research and Applications, 2020, pp. 1–7, https://doi.org/10.1145/3379157.3391414.

[29] J. Koskinen, R. Bednarik, H. Vrzakova, A.P. Elomaa, Combined gaze metrics as stress-sensitive indicators of microsurgical proficiency, Surg. Innovat. 27 (2020) 614–622, https://doi.org/10.1177/1553350620942980.

[30] M.F. Land, Vision, eye movements, and natural behavior, Vis. Neurosci. 26 (2009) 51–62, https://doi.org/10.1017/S0952523808080899.

[31] E.B. Lavoie, A.M. Valevicius, Q.A. Boser, O. Kovic, A.H. Vette, P.M. Pilarski, J. S. Hebert, C.S. Chapman, Using synchronized eye and motion tracking to determine high-precision eye-movement patterns during object-interaction tasks, J. Vis. 18 (2018) 18, https://doi.org/10.1167/18.6.18.

[32] B. Law, M.S. Atkins, A.E. Kirkpatrick, A.J. Lomax, Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment, in: Proceedings of the 2004 Symposium on Eye Tracking Research & Applications, 2004, pp. 41–48, https://doi.org/10.1145/968363.968370.

[33] D.R. Leff, D.R. James, F. Orihuela-Espina, K.W. Kwok, L.W. Sun, G. Mylonas, T. Athanasiou, A.W. Darzi, G.Z. Yang, The impact of expert visual guidance on trainee visual search strategy, visual attention and motor skills, Front. Hum. Neurosci. 9 (2015) 526, https://doi.org/10.3389/fnhum.2015.00526.

[34] T. Leppänen, H. Vrzakova, R. Bednarik, A. Kanervisto, A.P. Elomaa, A. Huotarinen, P. Bartczak, M. Fraunberg, J.E. Jääskeläinen, Augmenting microsurgical training: microsurgical instrument detection using convolutional neural networks, in: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2018, pp. 211–216, https://doi.org/10.1109/CBMS.2018.00044.

[35] J.K. Liu, V.N. Dodson, Combined petrosal approach for resection of recurrent brainstem cavernous malformation: operative video and technical nuances, Neurosurg. Focus: Video FOCVID 1 (2019) V18, https://doi.org/10.3171/2019.7.FocusVid.19229. (Accessed 12 March 2021).

[36] J.K. Liu, V.N. Dodson, Microsurgical resection of brainstem cervicomedullary ganglioglioma: operative video and technique of creating a surgical pseudoplane for near-total resection, Neurosurg. Focus: Video FOCVID 1 (2019) V13, https://doi.org/10.3171/2019.10.FocusVid.19413. (Accessed 12 March 2021).

[37] J.K. Liu, V.N. Dodson, Telovelar approach for microsurgical resection of fourth ventricular subependymoma arising from rhomboid fossa: operative video and technical nuances, Neurosurg. Focus: Video FOCVID 1 (2019) V5, https://doi.org/10.3171/2019.10.FocusVid.19452. (Accessed 12 March 2021).

[38] C. Loukas, Video content analysis of surgical procedures, Surg. Endosc. 32 (2018) 553–568, https://doi.org/10.1007/s00464-017-5878-1.

[39] H. Morisako, T. Goto, C.A. Bohoun, H. Arima, T. Ichinose, K. Ohata, Usefulness of the anterior transpetrosal approach for pontine cavernous malformations, Neurosurg. Focus: Video FOCVID 1 (2019), https://doi.org/10.3171/2019.7.FocusVid.19125. V4. (Accessed 12 March 2021).

[40] A. Nakazawa, K. Harada, M. Mitsuishi, P. Jannin, Real-time surgical needle detection using region-based convolutional neural networks, Int. J. Comput. Assist. Radiol. Surg. 15 (2020) 41–47, https://doi.org/10.1007/s11548-019-02050-9.

[41] J. Navarro, M. François, F. Mars, Obstacle avoidance under automated steering: impact on driving and gaze behaviours, Transport. Res. F Traffic Psychol. Behav. 43 (2016) 315–324, https://doi.org/10.1016/j.trf.2016.09.007.

[42] Neurosurgery Education and Training School, Department of Neurosurgery of the All India Institute of Medical Sciences, Brainstem Glioma Midbrain, 2021. https://www.youtube.com/watch?v=GraUJfBUR0s. (Accessed 12 March 2021).

[43] Neurosurgery Education and Training School, Department of Neurosurgery of the All India Institute of Medical Sciences, Focal Midbrain and Pons Glioma, 2021. https://www.youtube.com/watch?v=xCBelMhMSzc. (Accessed 12 March 2021).

[44] Neurosurgery Education and Training School, Department of Neurosurgery of the All India Institute of Medical Sciences, Hypothalamic Glioma Lamina Terminalis App, 2021. https://www.youtube.com/watch?v=4lCy-XviQWg. (Accessed 12 March 2021).

[45] C.I. Nwoye, D. Mutter, J. Marescaux, N. Padoy, Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos, Int. J. Comput. Assist. Radiol. Surg. 14 (2019) 1059–1067, https://doi.org/10.1007/s11548-019-01958-6.

[46] I. Pernek, A. Ferscha, A survey of context recognition in surgery, Med. Biol. Eng. Comput. 55 (2017) 1719–1734, https://doi.org/10.1007/s11517-017-1670-6.

[47] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2019. https://www.R-project.org/.

[48] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR), 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[49] C.E. Reiley, H.C. Lin, D.D. Yuh, G.D. Hager, Review of methods for objective surgical skill evaluation, Surg. Endosc. Other Intervent. Tech. 25 (2011) 356–366, https://doi.org/10.1007/s00464-010-1190-z.

[50] I. Rivas-Blanco, C.J. Perez-Del-Pulgar, I. Garcia-Morales, V.F. Munoz, A review on deep learning in minimally invasive surgery, IEEE Access 9 (2021) 48658–48678, https://doi.org/10.1109/ACCESS.2021.3068852.

[51] M. Sahu, A. Szengel, A. Mukhopadhyay, S. Zachow, Surgical phase recognition by learning phase transitions, Curr. Dir. Biomed. Eng. 6 (2020), https://doi.org/10.1515/cdbme-2020-0037.

[52] T. Schimmoeller, E.E. Neumann, T.F. Nagle, A. Erdemir, Reference tool kinematics-kinetics and tissue surface strain data during fundamental surgical acts, Sci. Data 7 (2020) 1–8, https://doi.org/10.1038/s41597-020-0359-0.

[53] S. Thiemjarus, A. James, G.Z. Yang, An eye–hand data fusion framework for pervasive sensing of surgical activities, Pattern Recogn. 45 (2012) 2855–2867, https://doi.org/10.1016/j.patcog.2012.01.008.
[54] G. Tien, M.S. Atkins, X. Jiang, B. Zheng, R. Bednarik, Verbal gaze instruction matches visual gaze guidance in laparoscopic skills training, in: Proceedings of the Symposium on Eye Tracking Research and Applications, 2014, pp. 331–334, https://doi.org/10.1145/2578153.2578217.
[55] T. Tien, P.H. Pucher, M.H. Sodergren, K. Sriskandarajah, G.Z. Yang, A. Darzi, Eye tracking for skills assessment and training: a systematic review, J. Surg. Res. 191 (2014) 169–178, https://doi.org/10.1016/j.jss.2014.04.032.
[56] Tzutalin, Labelimg. git code (2015), 2015. https://github.com/tzutalin/labelImg. (Accessed 14 September 2021).
[57] C.Y. Wang, H.Y.M. Liao, Y.H. Wu, P.Y. Chen, J.W. Hsieh, I.H. Yeh, Cspnet: a new backbone that can enhance learning capability of cnn, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 390–391, https://doi.org/10.1109/CVPRW50498.2020.00203.
[58] M. Wilson, J. McGrath, S. Vine, J. Brewer, D. Defriend, R. Masters, Psychomotor control in a virtual laparoscopic surgery training environment: gaze control parameters differentiate novices from experts, Surg. Endosc. 24 (2010) 2458–2464, https://doi.org/10.1007/s00464-010-0986-1.
[59] Y. Yamazaki, S. Kanaji, T. Matsuda, T. Oshikiri, T. Nakamura, S. Suzuki, Y. Hiasa, Y. Otake, Y. Sato, Y. Kakeji, Automated surgical instrument detection from laparoscopic gastrectomy video images using an open source convolutional neural network platform, J. Am. Coll. Surg. 230 (2020) 725–732, https://doi.org/10.1016/j.jamcollsurg.2020.01.037, e1.
[60] C. Yang, Z. Zhao, S. Hu, Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature, Comput. Assist. Surg. 25 (2020) 15–28, https://doi.org/10.1080/24699322.2020.1801842.
[61] Z. Zhao, T. Cai, F. Chang, X. Cheng, Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade, Healthc. Technol. Lett. 6 (2019) 275–279, https://doi.org/10.1049/htl.2019.0064.
[62] Z. Zhao, S. Voros, Z. Chen, X. Cheng, Surgical tool tracking based on two cnns: from coarse to fine, J. Eng. (2019) 467–472, https://doi.org/10.1049/joe.2018.9401, 2019.
[63] M. Zhou, M. Hamad, J. Weiss, A. Eslami, K. Huang, M. Maier, C.P. Lohmann, N. Navab, A. Knoll, M.A. Nasseri, Towards robotic eye surgery: marker-free, online hand-eye calibration using optical coherence tomography images, IEEE Robot. Autom. Lett. 3 (2018) 3944–3951, https://doi.org/10.1109/LRA.2018.2858744.