CATARACTS: Challenge on Automatic Tool Annotation for cataRACT Surgery

Hassan Al Hajj[a], Mathieu Lamard[b,a], Pierre-Henri Conze[c,a], Soumali Roychowdhury[d], Xiaowei Hu[e], Gabija Maršalkaitė[f], Odysseas Zisimopoulos[g], Muneer Ahmad Dedmari[i], Fenqiang Zhao[k], Jonas Prellberg[l], Manish Sahu[m], Adrian Galdran[p], Teresa Araújo[o,p], Duc My Vo[q], Chandan Panda[r], Navdeep Dahiya[s], Satoshi Kondo[t], Zhengbing Bian[d], Arash Vahdat[d], Jonas Bialopetravičius[f], Evangello Flouty[g], Chenhui Qiu[k], Sabrina Dill[m], Anirban Mukhopadhyay[n], Pedro Costa[p], Guilherme Aresta[o,p], Senthil Ramamurthy[s], Sang-Woong Lee[q], Aurélio Campilho[o,p], Stefan Zachow[m], Shunren Xia[k], Sailesh Conjeti[i,j], Danail Stoyanov[g,h], Jogundas Armaitis[f], Pheng-Ann Heng[e], William G. Macready[d], Béatrice Cochener[b,a,u], Gwenolé Quellec[a,*]

[a] *Inserm, UMR 1101, Brest, F-29200 France*
[b] *Univ Bretagne Occidentale, Brest, F-29200 France*
[c] *IMT Atlantique, LaTIM UMR 1101, UBL, Brest, F-29200 France*
[d] *D-Wave Systems Inc., Burnaby, BC, V5G 4M9 Canada*
[e] *Dept. of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China*
[f] *Oxipit, UAB, Vilnius, LT-10224 Lithuania*
[g] *Digital Surgery Ltd, EC1V 2QY, London, UK*
[h] *University College London, Gower Street, WC1E 6BT, London, UK*
[i] *Chair for Computer Aided Medical Procedures, Faculty of Informatics, Technical University of Munich, Garching b. Munich, 85748 Germany*
[j] *German Center for Neurodegenrative Diseases (DZNE), Bonn, 53127 Germany*
[k] *Key Laboratory of Biomedical Engineering of Ministry of Education, Zhejiang University, HangZhou, 310000 China*
[l] *Dept. of Informatics, Carl von Ossietzky University, Oldenburg, 26129 Germany*
[m] *Department of Visual Data Analysis, Zuse Institute Berlin, Berlin, 14195 Germany*
[n] *Department of Computer Science, Technische Universität Darmstadt, 64283 Darmstadt, Germany*
[o] *Faculdade de Engenharia, Universidade do Porto, Porto, 4200-465 Portugal*
[p] *INESC TEC - Instituto de Engenharia de Sistemas e Computadores - Tecnologia e Ciência, Porto, 4200-465 Portugal*
[q] *Gachon University, 1342 Seongnamdaero, Sujeonggu, Seongnam 13120, Korea*
[r] *Epsilon, Bengaluru, Karnataka 560045, India*
[s] *Laboratory of Computational Computer Vision, Georgia Tech, Atlanta, GA 30332, USA*
[t] *Konica Minolta, Inc., Osaka, 569-8503 Japan*
[u] *Service d'Ophtalmologie, CHRU Brest, Brest, F-29200 France*

Abstract

Surgical tool detection is attracting increasing attention from the medical image analysis community. The goal generally is not to precisely locate tools in images, but rather to indicate which tools are being used by the surgeon at each instant. The main motivation for annotating tool usage is to design efficient solutions for surgical workflow analysis, with potential applications in report generation, surgical training and even real-time decision support. Most existing tool annotation algorithms focus on laparoscopic surgeries. However, with 19 million interventions per year, the most common surgical procedure in the world is cataract surgery. The CATARACTS challenge was organized in 2017 to evaluate tool annotation algorithms in the specific context of cataract surgery. It relies on more than nine hours of videos, from 50 cataract surgeries, in which the presence of 21 surgical tools was manually annotated by two experts. With 14 participating teams, this challenge can be considered a success. As might be expected, the submitted solutions are based on deep learning. This paper thoroughly evaluates these solutions: in particular, the quality of their annotations are compared to that of human interpretations. Next, lessons learnt from the differential analysis of these solutions are discussed. We expect that they will guide the design of efficient surgery monitoring tools in the near future.

Keywords: cataract surgery, video analysis, deep learning, challenge

1. Introduction

Video recording is a unique solution to collect information about a surgery. Combined with computer vision and machine learning, it allows a wide range of applications, including automatic report generation, surgical skill evaluation and training, surgical workflow optimization, as well as warning and recommendation generation. Key indicators of what the surgeon is doing at any given time are the surgical tools that he or she is using. Therefore, several tool detection techniques have been presented in recent years (Bouget et al., 2017). The Challenge on Automatic Tool Annotation for cataRACT Surgery (CATARACTS)[1] was organized in 2017 to evaluate the relevance of these techniques and novel ones in the context of

[*] LaTIM - IBRBS - CHRU Morvan - 12, Av. Foch
29609 Brest CEDEX - FRANCE
Tel.: +33 2 98 01 81 29 / Fax: +33 2 98 01 81 24
Email address: gwenole.quellec@inserm.fr (Gwenolé Quellec)

[1] https://cataracts.grand-challenge.org

cataract surgery. This paper introduces the results and main conclusions of the CATARACTS challenge.

A cataract is an opacification of the crystalline lens, a biconvex eye structure located behind the iris. Normally transparent, this lens helps to focus light onto the retina and provides accommodation. Cataract develops with aging, general disease, congenital disorder or injury, and leads to a decrease in vision. Symptoms include cloudy or blurred vision, faded colors, glare, poor night vision and double vision. This is the most common cause of vision loss and blindness in the world: according to the World Health Organization, the number of cataract blind people will reach 40 million in 2025 (Wang et al., 2016). When vision loss interferes with everyday activities, cataract surgery is recommended (Kessel et al., 2016). This is the most frequently performed surgical procedure in many economically developed countries (Erie, 2014; Wang et al., 2016). Its purpose is to remove the crystalline lens and replace it with an artificial intraocular lens (IOL). Physiologically, the crystalline lens is contained in a bag, which is connected to the ciliary body by a zonule. Until the early 1960s, the lens was removed with its bag in a so called "intracapsular" extraction, using cryoextraction for a better hold of the lens (Krwawicz, 1961): this required a 180 degree incision around the cornea (Olson, 2018). Later, it was replaced with "extracapsular" lens removal: the capsular bag is left inside the eye, allowing the IOL to be implanted in it. The advent of phacoemulsification definitely revolutionized the surgery in terms of safety, efficacy and reproducibility. Thanks to an ultrasonic handpiece, the crystalline lens is fragmented into small pieces, which can be removed by suction through a small incision (Kelman, 1967). Introduced in 1967, this technique started emerging in routine practice in the 1980s and improved over time to require less ultrasound energy and smaller incision size (about 1.8 to 2.2 millimeters today). The IOL, initially made of rigid polymethylmethacrylate, also required tremendous evolution in biomaterials to allow insertion through a small incision, hence the development of foldable IOLs made of silicone and then of hydrophilic or hydrophobic acrylics (Seward, 1997). The result of a smaller incision is less astigmatism with faster recovery and decrease of postoperative complications (Riaz et al., 2006). Recent technological advances include femtosecond laser-assisted surgery, which automates the process of crystalline lens fragmentation (Popovic et al., 2016), and premium IOLs: toric optics for astigmatism correction, multifocal and extended depth of focus IOLs for presbyopia correction (de Silva et al., 2016; Cochener et al., 2018).

Because of its frequency, cataract surgery is the first surgery that eye surgeons need to master (Kaplowitz et al., 2018 Mar-Apr): this is one major motivation for developing computer-aided decision tools for cataract surgery. One way to help surgeons during their training period is to analyze their surgical workflow (Charrière et al., 2017). Through comparisons with more experienced surgeons, postoperatively, it may help self-evaluation. Surgical workflow analysis is also useful for workflow optimization: by analyzing the workflow of several surgeries, and their outcome, lessons can be learnt and best practices can be identified. The same principle can be applied intraoperatively: warnings and recommendations can be generated automatically whenever an unusual or suboptimal workflow pattern is detected. Surgical workflow optimization and recommendation generation can be useful for cataract surgery, after a new technological evolution or for training. It is probably even more useful for rare surgeries, where training data is more difficult to collect. So instead of analyzing the surgical workflow directly, which is specific to each surgery, we propose to focus on tool usage analysis instead: because tools used in cataract surgeries are similar to those used in other surgeries, tool usage annotation algorithms can be easily transferred to other surgeries.

In recent years, the number of medical image analysis challenges has exploded. According to Grand-Challenge[2], which lists those challenges and hosts some of them, two challenges were organized per year in 2007 and 2008; their number progressively increased to 15 per year in 2012 and 2013; more than 20 challenges are now organized every year. The first challenge organized in the context of ophthalmology was the Retinopathy Online Challenge in 2009 (Niemeijer et al., 2010): the goal was to detect signs of diabetic retinopathy in fundus photographs. Two other challenges were organized on the same topic: the Diabetic Retinopathy Detection challenge in 2015[3] and the IDRiD challenge in 2018.[4] The detection and segmentation of retinal anomalies in optical coherence tomography images was the topic of three other challenges: the Retinal Cyst Segmentation Challenge in 2015,[5] RETOUCH[6] and ROCC[7] in 2017. However, CATARACTS is the only challenge related to ophthalmic surgery and ophthalmic video analysis. Outside the scope of ophthalmology, three other challenges about surgery video analysis have been organized: EndoVis in 2015 and 2017 (Bernal et al., 2017),[8] and M2CAI in 2016 (Twinanda et al., 2016).[9] Although those three challenges are related to digestive surgery, they share similarities with CATARACTS. In particular, M2CAI had a sub-challenge on tool detection and both editions of EndoVis had a sub-challenge on tool segmentation. What makes tool detection particularly challenging in CATARACTS, compared to EndoVis and M2CAI, probably is the large range of tools that must be recognized. The reason is that digestive surgeries addressed in EndoVis and M2CAI rely on robotic arms with a standardized set of tools, whereas eye surgeons operate manually and can therefore chose from a wide selection of tools from several manufacturers.

The remainder of the paper is organized as follows. Section 2 reviews the recent literature about surgical tool analysis. The setup of the CATARACTS challenge is described in section 3. Competing solutions are presented in section 4. Results are reported in section 5. The paper ends with a discussion and conclusions in section 6.

[2] https://grand-challenge.org/All_Challenges

[3] http://www.kaggle.com/c/diabetic-retinopathy-detection

[4] https://idrid.grand-challenge.org

[5] https://optima.meduniwien.ac.at/research/challenges

[6] https://retouch.grand-challenge.org

[7] https://rocc.grand-challenge.org

[8] https://endovis.grand-challenge.org

[9] http://camma.u-strasbg.fr/m2cai2016

2. Review of Surgical Tool Analysis

Over the past decades, surgical tool analysis mostly relied on external markers attached to the tools. This includes shape markers (Casals et al., 1996), color markers (Ko et al., 2005), optical markers (Krupa et al., 2003), acoustic markers (Chmarra et al., 2007) and RFID systems (Miyawaki et al., 2009). With the progress of computer vision, solutions for vision-based and marker-less tool analysis have emerged. Bouget et al. (2017) thoroughly reviewed the literature of this domain until 2015; recent trends are discussed hereafter.

2.1. Clinical Applications

In terms of applications, it should be noted that most solutions were developed to monitor endoscopic videos for minimally invasive surgeries, with or without robotic assistance (Sarikaya et al., 2017; Ross et al., 2018; Wesierski and Jezierska, 2018; Du et al., 2018; Allan et al., 2018). Other imaging modalities include:

- microscopy, for neurosurgery (Leppänen et al., 2018), retinal surgery (Alsheakhali et al., 2016b; Rieke et al., 2016a; Kurmann et al., 2017; Laina et al., 2017) and cataract surgery (Al Hajj et al., 2017a),
- OCT, (Gessert et al., 2018), for ophthalmic microsurgery (Zhou et al., 2017; Keller et al., 2018),
- X-rays, for endovascular surgery (Chang et al., 2016) and face surgery (Kügler et al., 2018),
- ultrasound (Rathinam et al., 2017), for intraplacental interventions (García-Peraza-Herrera et al., 2016),
- and RBGD, for orthopedic surgery (Lee et al., 2017b).

2.2. Computer Vision Tasks

In terms of computer vision tasks, multiple problems have been addressed in the recent literature. These tasks can be categorized according the precision of the desired outputs. The finest task is tool segmentation (Bodenstedt et al., 2016 Feb-Mar; García-Peraza-Herrera et al., 2016, 2017; Attia et al., 2017; Lee et al., 2017b; Zhou et al., 2017; Ross et al., 2018; Su et al., 2018). This includes multi-label tool segmentation for articulated tools (Laina et al., 2017): each tool part is associated with one label. A coarser task is tool detection or localization (Chang et al., 2016; Leppänen et al., 2018): the goal typically is to detect the tool tip (Furtado et al., 2016; Chen et al., 2017; Czajkowska et al., 2018) or the tool edges (Agustinos and Voros, 2015; Chen et al., 2017). For articulated instruments, the goal is also to detect the tool parts (Wesierski and Jezierska, 2017 Aug-Sep, 2018) or the articulations between them (Laina et al., 2017; Du et al., 2018). For flexible instruments, the goal is also to detect the tool centerline (Chang et al., 2016). Tool detection generally is an intermediate step for tool tracking, the process of monitoring tool location over time (Du et al., 2016; Rieke et al., 2016a; Lee et al., 2017b; Zhao et al., 2017; Czajkowska et al., 2018; Ryu et al., 2018; Keller et al., 2018), and pose estimation, the process of inferring a 2-D pose (Rieke et al., 2016b; Kurmann et al., 2017; Alsheakhali et al., 2016b; Du et al., 2018; Wesierski and Jezierska, 2018) or a 3-D pose (Allan et al., 2018; Gessert et al., 2018) based on the location of tool elements. Tasks associated with tool detection also include velocity estimation (Marban et al., 2017) and instrument state recognition (Sahu et al., 2016a). All the above tasks are directly useful to the surgeon: they can be used for improved visualization, through augmented or mixed reality (Frikha et al., 2016 Nov-Dec; Bodenstedt et al., 2016 Feb-Mar; Lee et al., 2017b,a).

Finally, the coarsest task is tool presence detection: the goal is to determine which tools are present or active in each frame of the surgical video (Sahu et al., 2017; Primus et al., 2016; Hu et al., 2017; Sarikaya et al., 2017; Twinanda et al., 2017; Wang et al., 2017; Al Hajj et al., 2017a; Jin et al., 2018). This is the task addressed in this paper. Unlike finer tasks, the usefulness of this task is indirect: it is mainly used to analyze the surgical workflow (Twinanda et al., 2017).

2.3. Computer Vision Algorithms

Various computer vision algorithms have been proposed to address these tasks. Until early 2017, tool detection relied heavily on handcrafted features, including Gabor filters (Czajkowska et al., 2018), Frangi filters (Agustinos and Voros, 2015; Chang et al., 2016), color-based features (Primus et al., 2016; Rieke et al., 2016a), histograms of oriented gradients (Rieke et al., 2016a; Czajkowska et al., 2018), SIFT features (Du et al., 2016), ORB features (Primus et al., 2016) and local binary patterns (Sahu et al., 2016a). For tool segmentation, similar features have been extracted within superpixels (Bodenstedt et al., 2016 Feb-Mar). These features were processed either by a machine learning algorithm, such as a support vector machine (Primus et al., 2016; Wesierski and Jezierska, 2018), a random forest (Bodenstedt et al., 2016 Feb-Mar; Rieke et al., 2016a,b) or AdaBoost (Sahu et al., 2016a), or by a parametric model, such as a generalized Hough transform (Du et al., 2016; Frikha et al., 2016 Nov-Dec; Czajkowska et al., 2018) or a B-spline model (Chang et al., 2016). Note that template matching techniques have also been used to deal with articulated instruments (Ye et al., 2016; Wesierski and Jezierska, 2018).

Since 2017, most tool analysis solutions rely on deep learning. For tool detection, convolutional neural networks (CNNs) were used to recognize images patches containing tool pixels (Alsheakhali et al., 2016a; Chen et al., 2017; Zhao et al., 2017). The use of region proposal networks was also investigated (Sarikaya et al., 2017; Jin et al., 2018). Several CNN architectures were experimented for tool segmentation: fully convolutional networks (García-Peraza-Herrera et al., 2016; Zhou et al., 2017), U-net (Ross et al., 2018) or custom encoder/decoder CNN architectures (García-Peraza-Herrera et al., 2017; Attia et al., 2017; Laina et al., 2017). The use of generative adversarial networks was proposed to train or pre-train segmentation CNNs: a tool segmentation CNN (Ross et al., 2018) and a specular highlight segmentation and removal CNN (Funke et al., 2018). For pose estimation, regression CNNs were proposed (Du et al., 2018; Gessert et al., 2018; Kügler et al., 2018), which eliminates the need to explicitly localize tools as an intermediate step. Note that multi-task CNNs have been designed: Laina et al. (2017) jointly segments the tools and detects the joints between tool parts, Kurmann et al. (2017) jointly recognizes the tools and detects the joints between tool parts, Du et al. (2018) jointly detects the joints between tool parts and estimates 2-D poses, Jin et al. (2018) and Hu et al. (2017) jointly determines tool presence and tool localization. To take time information into account, proposed solutions sometimes took advantage of the optical flow (Czajkowska et al., 2018) or relied on temporal filtering techniques, such as a Kalman filter (Ryu et al., 2018) or a recurrent neural network (RNN) (Attia et al., 2017; Marban et al., 2017). This is typically useful for tool tracking (Chen et al., 2017; Marban et al., 2017; Ryu et al., 2018; Czajkowska et al., 2018), but it was also used to speed up tool segmentation (García-Peraza-Herrera et al., 2016) or to improve tool presence detection (Al Hajj et al., 2017a).

2.4. Tool Presence Detection Pipeline

Nowadays, tool presence detection algorithms also rely on CNNs (Al Hajj et al., 2017b; Hu et al., 2017; Kurmann et al., 2017; Sahu et al., 2017; Twinanda et al., 2017) or CNN ensembles (Wang et al., 2017). These CNNs accept full video frames as input and compute a probability of presence for each surgical tool in the input frame. These CNNs are generally trained through transfer learning (Yosinski et al., 2014; Litjens et al., 2017): image classification models, typically pre-trained on ImageNet[10], are fine-tuned on individual frames extracted from training videos. This strategy was followed by the winners of the M2CAI tool detection sub-challenge (Raju et al., 2016; Sahu et al., 2016b; Twinanda et al., 2017; Zia et al., 2016). Once CNNs are trained, their predictions can be improved using a temporal model. In the simplest scenario, each prediction signal are smoothed by a usual temporal filter (e.g. a median filter) to compensate for short-term occlusion or image quality problems. Whenever long-term relationships between events are important, a RNN can be used instead (Yao et al., 2015; Donahue et al., 2017). CNN+RNN models have thus been used for surgical workflow analysis in endoscopy videos (Twinanda et al., 2017; Jin et al., 2016; Bodenstedt et al., 2017). Given the correlation between surgical workflow and tool usage, such an approach also seems relevant for tool usage annotation in surgery videos (Mishra et al., 2017; Al Hajj et al., 2018).

3. Challenge Description

3.1. Video Collection

The challenge relies on a dataset of 50 videos of phacoemulsification cataract surgeries performed in Brest University Hospital between January 22, 2015 and September 10, 2015. Reasons for surgery included age-related cataract, traumatic cataract and refractive errors. Patients were 61 years old on average (minimum: 23, maximum: 83, standard deviation: 10). There were 38 females and 12 males. Informed consent was obtained from all patients. Surgeries were performed by three surgeons: a renowned expert (48 surgeries), a one-year experienced surgeon (1 surgery) and an intern (1 surgery). Surgeries were performed under an OPMI Lumera T microscope (Carl Zeiss Meditec, Jena, Germany). Videos were recorded with a 180I camera (Toshiba, Tokyo, Japan) and a MediCap USB200 recorder (MediCapture, Plymouth Meeting, USA). The frame definition was 1920x1080 pixels and the frame rate was approximately 30 frames per second. Videos had a duration of 10 minutes and 56 s on average (minimum: 6 minutes 23 s, maximum: 40 minutes 34 s, standard deviation: 6 minutes 5 s). In total, more than nine hours of surgery have been video recorded.

3.2. Training and Test Set Separation

The dataset was divided evenly into a training set (25 videos) and a test set (25 videos). Division was made in such a way that each tool appears in the same number of videos from both sets (plus or minus one). No validation dataset was provided: participants were given the responsibility to divide the training set into a learning subset and a validation subset. Ground truth was collected similarly for training and test videos, as described hereafter.

3.3. Tool Usage Annotation

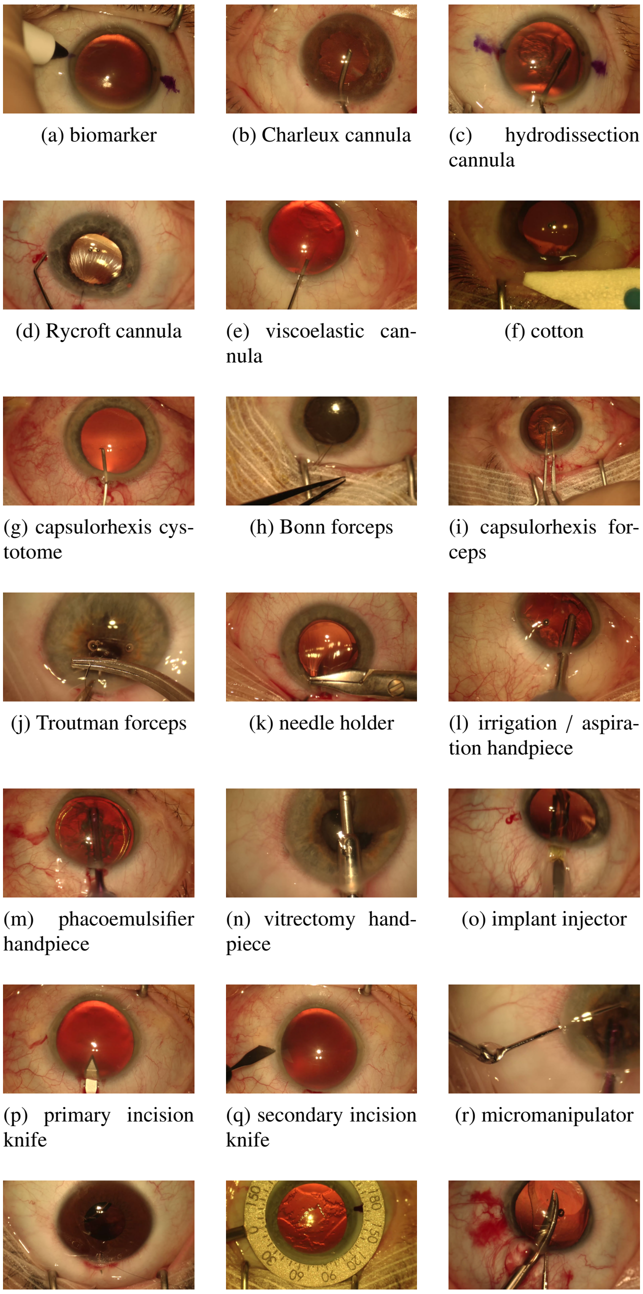
All surgical tools visible in microscope videos were first enumerated and labeled by the surgeons: a list of 21 tools was obtained (see Fig 1). Then, the usage of each tool in videos was annotated independently by two non-clinical experts. A tool was considered to be in use whenever it was in contact with the eyeball. Therefore, a timestamp was recorded by both experts whenever one tool came into contact with the eyeball, and also when it stopped touching the eyeball. Up to three tools may be used simultaneously: two by the surgeon (one per hand) and sometimes one by an assistant. Annotations were performed at the frame level, using a web interface connected to an SQL database. Finally, annotations from both experts were adjudicated: whenever expert 1 annotated that tool A was being used, while expert 2 annotated that tool B was being used instead of A, experts watched the video together and jointly determined the actual tool usage. However, the precise timing of tool/eyeball contacts was not adjudicated. Therefore, a probabilistic reference standard was obtained:

- 0: both experts agree that the tool is not being used,
- 1: both experts agree that the tool is being used,
- 0.5: experts disagree.

Inter-rater agreement, before and after adjudication, is reported in Table 1. A chord diagram[11] illustrating the co-occurrence of tools in training video frames is reported in Fig. 2.

[10] www.image-net.org

[11] http://mkweb.bcgsc.ca/tableviewer/



(a) biomarker (b) Charleux cannula (c) hydrodissection cannula

(d) Rycroft cannula (e) viscoelastic cannula (f) cotton

(g) capsulorhexis cystotome (h) Bonn forceps (i) capsulorhexis forceps

(j) Troutman forceps (k) needle holder (l) irrigation / aspiration handpiece

(m) phacoemulsifier handpiece (n) vitrectomy handpiece (o) implant injector

(p) primary incision knife (q) secondary incision knife (r) micromanipulator

(s) suture needle (t) Mendez ring (u) Vannas scissors

Figure 1: Surgical tools annotated in videos

Tool	Agreement before adjudication	Agreement after adjudication	% of training frames in use
biomarker	0.835	0.835	0.0168 %
Charleux cannula	0.949	0.963	1.79 %
hydrodissection cannula	**0.868**	**0.982**	2.43 %
Rycroft cannula	**0.882**	**0.919**	3.18 %
viscoelastic cannula	**0.860**	**0.975**	2.54 %
cotton	0.947	0.947	0.751 %
capsulorhexis cystotome	0.994	0.995	4.42 %
Bonn forceps	0.793	0.798	1.10 %
capsulorhexis forceps	0.836	0.849	1.62 %
Troutman forceps	0.764	0.764	0.258 %
needle holder	0.630	0.630	0.0817 %
irrigation/aspiration handpiece	0.995	0.995	14.2%
phacoemulsifier handpiece	0.996	0.997	15.3 %
vitrectomy handpiece	0.998	0.998	2.76 %
implant injector	0.980	0.980	1.41 %
primary incision knife	0.959	0.961	0.700 %
secondary incision knife	0.846	0.852	0.522 %
micromanipulator	0.990	0.995	17.6 %
suture needle	0.893	0.893	0.219 %
Mendez ring	0.941	0.953	0.100 %
Vannas scissors	0.823	0.823	0.0443 %

Table 1: Statistics about tool usage annotation in the CATARACTS dataset. The first two columns indicate inter-rater agreement (Cohen's kappa) before and after adjudication; the largest changes are in bold. The last column indicates the prevalence of each tool in the training subset, ignoring the frames where experts disagree about the usage of that tool, even after adjudication.

3.4. *Performance Evaluation of a Submission*

Tool usage predictions submitted by a participant for test videos were evaluated as follows. A figure of merit was first computed for each tool label T: the annotation performance for tool T was defined as the area $A(T)$ under the receiver-operating characteristic (ROC) curve (see Fig. 5). This curve was obtained by varying a cutoff on the confidence level for tool T provided by the participant for each frame in the test set. Frames associated with a disagreement between experts (reference standard = 0.5 for tool T) were ignored when computing the ROC curve. Then, a global figure of merit was defined: it was simply defined as the mean $A(T)$ value over all tool labels T. The evaluation script was made publicly available at the beginning of the challenge.

3.5. *Rules of the Challenge*

Training videos, with their tool usage annotations, as well as test videos, without their annotations, were released on April 1, 2017. The challenge has been open for submissions during eight months, from April 1, 2017 to November 30, 2017. In order to stimulate competition and to explore more solutions, participants were allowed to submit multiple solutions throughout this period. However, two restrictions were imposed on re-submissions:

1. Each submission was required to be substantially different from the previous ones. Typically, a first submission may consist of a CNN only, a second one may consist of an ensemble of CNNs, and third one may include a temporal sequencer. However, submitting the same algorithm with different meta-parameters was not allowed. This rule was fixed to minimize the risk of influencing the solution's behavior with test data. To allow verification of this rule by the organizers, a technical report was required for each submission and re-submission.

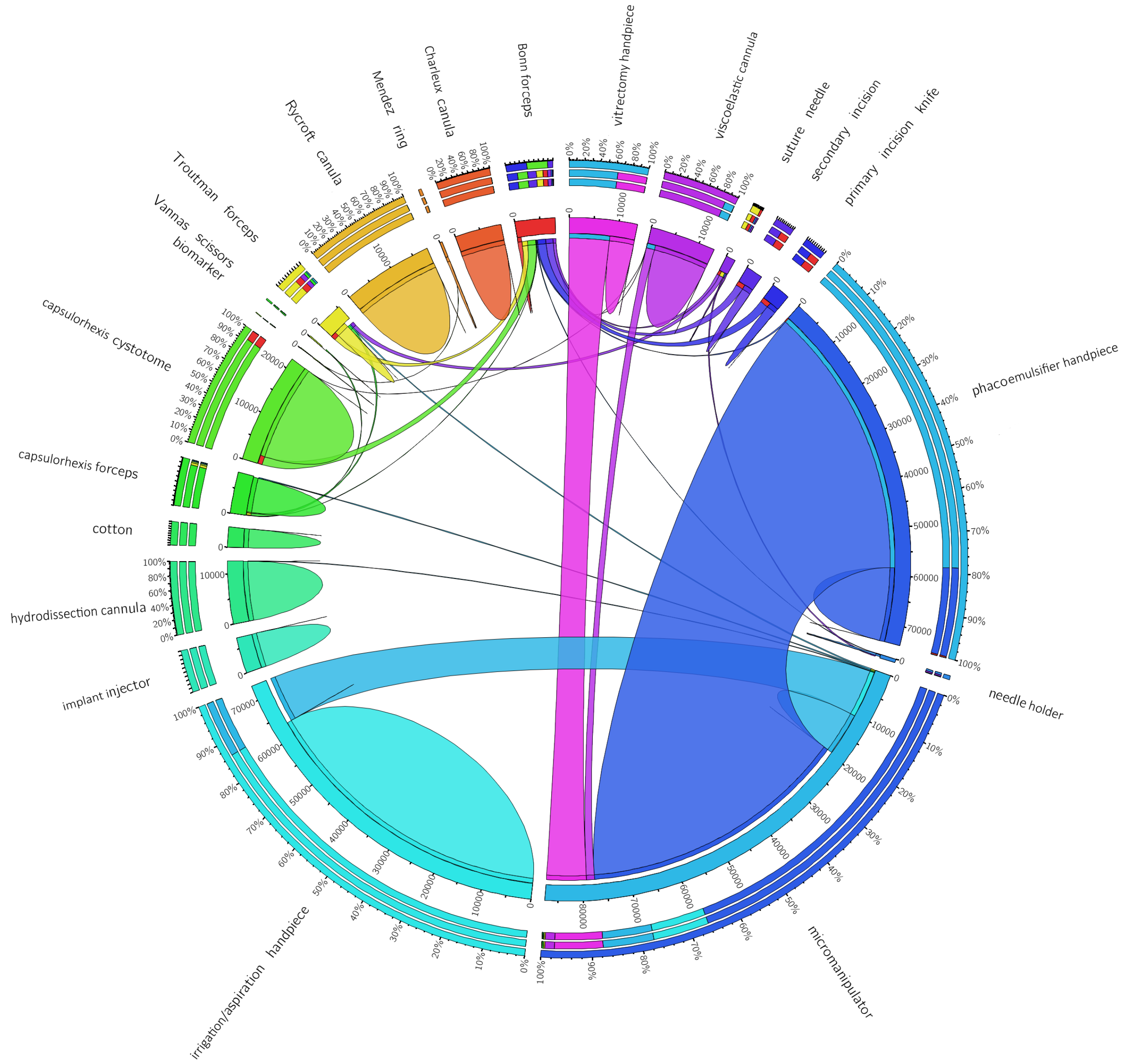


Figure 2: Chord diagram illustrating tool co-occurrence in training video frames. This figure shows, for instance, that the phacoemulsifier handpiece is used in 74,000 frames and that, in 78,5% of these frames, it is used in conjunction with the micromanipulator.

2. Technical reports and performance scores were immediately published on the challenge website and no resubmission was evaluated for a week. This rule was fixed to balance the inequities between teams submitting multiple solutions and those submitting only once: the latter can benefit from experience gained by the former.

For each team, the solution with maximal performance among all submissions (if more than one) was retained to compile the final team ranking. Two submissions were excluded from the establishment of this ranking by virtue of the one week waiting rule: the scheduled evaluation date occurred after the challenge closing date. However, they are discussed in the following section anyway. Solutions submitted by the organizers (LaTIM) are not included in the team ranking, but are also discussed in this paper.

4. Competing Solutions

Fourteen teams competed in this challenge. Their solutions, as well as the organizers' solution, are described hereafter. To allow comparisons between these solutions, key elements are reported in Tables 2, 3, 4, 5 and 6.

4.1. *VGG fine-tuning*

The VGG fine-tuning solution uses a CNN with weights pretrained on the ImageNet dataset. The base network is VGG-16 (Simonyan and Zisserman, 2015). The last fully connected layer, namely ‘fc8’, was changed to have twenty-one output neurons, each representing the likelihood that one tool is being used by the surgeon in the input image. The last two fully connected layers, namely ‘fc7’ and ‘fc8’, were fine-tuned using the CATARACTS training dataset. The CNN processes images with 288×288 pixels. It was trained using a stochastic gradient descent with a learning rate of 0.001 and a momentum of 0.9. The mini-batch size was set to 48 and the number of epochs to 80. A weighted loss function was used: a weight of one was assigned to label 0 (tool not being used) and a weight of thirty was assigned to label 1 (tool in use). No random distortions are applied to input images during training and inference.

4.2. LCCV-Cataract

The LCCV-Cataract solution relies on an Inception-v3 CNN (Szegedy et al., 2016) pre-trained on ImageNet. The major difference with other solutions is that a multi-class classifier was trained *(each image has exactly one label)*, rather than a multi-label classifier *(each image may have zero, one or multiple labels)*. Twenty-two mutually exclusive classes were defined: each of the first 21 classes predicts the usage of one tool and the 22$^{\text{nd}}$ class predicts the absence of tool usage. For compatibility reasons, all video frames associated with multiple tools in the CATARACTS dataset were ignored during training. The CNN processes images with 299×299 pixels. It was fine-tuned with a learning rate of 0.01 for several thousand iterations with cross-entropy loss. During inference, the purpose of the 22$^{\text{nd}}$ class it to lower the probability of the other 21 classes when no tool appears to be in use. No random distortions are applied to input images during training and inference.

4.3. AUGSQZNT

The AUGSQZNT solution extends SqueezeNet, a lightweight CNN (Iandola et al., 2016) with weights pre-trained on ImageNet. The proposed architecture starts with three blocks of convolutional layers and then splits into three parts: one part for the ‘cannula’ set of labels, one part for the ‘forceps’ set and one part for the rest. The ‘forceps’ split of the network uses softmax activations while the other two use sigmoid activations. For validation, 5 complete videos and selected frames containing approximately 20% of frames labelled biomarker, needle holder, vitrectomy handpiece and Vannas scissors from 3 videos were kept aside from training. This was to ensure that each label has approximately 15-20% representation in the validation set. The frames were extracted at 10 frames per second although for rare classes, the frames were duplicated up to 50 times after extraction. Afterwards, all frames were augmented using vertical and horizontal flipping and randomly cropping 70%. The CNN was trained using a binary cross entropy loss function with a 80:10:10 weight ratio assigned to each network split. The Adam optimizer (Kingma and Ba, 2015) was used with a learning schedule starting with the learning rate of 0.01 and subsequently dividing by 10 after every 3 epochs with no improvement in validation loss. During inference, 5-fold test time augmentation is performed by taking the center, top left, top right, bottom right and bottom left patches from each frame in the test dataset. The predictions are averaged across the 5 patches for each frame.

4.4. SurgiToolNet

The SurgiToolNet solution is a deep learning network based on DenseNet-161 (Huang et al., 2017). The DenseNet-161 model was pre-trained on ImageNet to accelerate the training process. To use the DenseNet-161 network as a multi-label classifier, a Euclidean loss layer was plugged into the end of the network to compute the sum of squares of differences between the predicted output and the ground truth input. The CNN processes images with 224×224 pixels. It was fine-tuned using stochastic gradient descent with a momentum of 0.9. The initial learning rate was set to 0.001, and was divided by 10 after 50,000 iterations. In the deployment process, a binary classification layer was added at the end of this network: this layer is used to threshold the outputs of the fully connected layer and classify them into binary labels $\in \{0, 1\}$, indicating whether or not each tool is being used by the surgeon in the current frame.

4.5. CRACKER

CRACKER uses a frame-wise tool detector, based on a ResNet-34 (He et al., 2016) pre-trained on ImageNet, followed by field knowledge-based temporal filtering. The optimizer is the SGDR (Loshchilov and Hutter, 2017) and the loss function is the categorical cross entropy log loss.

Frame-wise tool detector: The model was fine-tuned with a 1:2 subsample of the CATARACTS dataset rescaled to 128×128 pixels. First, the top of the network was trained for a fixed number of epochs. Then, the learning rate was reduced by 1/3 at each 1/3 of the network depth. Finally, the entire network was trained until the cross entropy log loss stagnated in the validation set. Test predictions are the result of the average of the model’s output over 4 different test-time augmented versions of the frames.

Knowledge-based temporal filtering: First, the temporally sorted predictions are median-filtered with a sliding filter of size 11. For the irrigation/aspiration handpiece, phacoemulsifier handpiece and implant injector, the filter size was set to 101 instead. All signals are then processed based on the surgical procedure: 1) the irrigation/aspiration and vitrectomy handpieces (IA, V, respectively) usually proceed the phacoemulsifier because the latter is used for lens destruction; 2) the implant injector can never come before IA or V pieces since the implant can only be injected into the eye once the damaged lens has been removed and 3) the Rycroft cannula should not come before IA or V since it is used for refilling the lens in the end of the surgery. With that in mind, the first occurrence of $\text{probability}_{IA} > 0.5$ or $\text{probability}_{V} > 0.5$ is used for zeroing erroneous predictions of the above-mentioned tools.

4.6. MIL+resnet

The main contribution of the MIL+resnet solution is the decoupling of the initial task into a binary tool detection stage followed by a 21-class classification to determine the tools present on each given frame. The binary tool detection model is based on the Multiple-Instance Learning (MIL) framework (Quellec et al., 2017a). The MIL assumption was interpreted in this context as follows: image patches are considered as instances, a patch containing (part of) a tool is considered as a positive instance, and a patch with no signs of tool presence is considered as a negative instance. Accordingly, a given image is considered as a bag containing instances. The sole presence of a positive instance is enough to declare the associated bag as containing a tool, whereas in order for a frame/bag to be declared as not containing tools, it must be composed only of negative instances.

In this stage, a standard CNN architecture was employed, namely the Inception-v3 network, with initial weights pre-trained on the ImageNet dataset. In order to deal with patches, the architecture was modified to perform patch-level classification given the full input image. The deeper layers of the Inception-v3 network were discarded, since the receptive field of each layer grows as the network gets deeper. By discarding deeper layers of the network, the receptive field of the output layer can be effectively reduced. The predicted patch labels must then be combined to produce an image-level prediction. In order to follow the standard MIL assumption, patch predictions are merged into a single prediction by means of a max-pooling function.

The binary tool detector was trained on a binarized tool/no-tool version of the provided ground-truth. The resulting model was applied on the test set to retain frames that contained tools. The predictions on test set were temporally smoothed with a trimmed mean filter to add some robustness. Afterwards, a ResNet CNN was trained only on tool-containing frames, in order to learn to classify which were the present tools. This second stage was considered as a standard 21-class multi-label classification problem. Finally, the trained model was applied only to test frames that had been predicted as containing tools to decide which tools were present at each moment on the videos from the test set.

4.7. ZIB-Res-TS

The framework of the ZIB-Res-TS comprises of three main parts: stratification of the data, a classification model and temporal smoothing as a post-processing step. Since multiple tools can be visible in an image and tool co-occurrence frequency varies within the dataset, label-set sampling (Sahu et al., 2017) was applied to the data to reduce the bias caused by highly frequent tool co-occurrences. This approach relies on stratified sampling based on the co-occurrences of tools as disjoint classes. The model consists of ResNet-50 which was pre-trained on ImageNet and fine-tuned on the CATARACTS dataset by adding a global average pooling and a fully connected layer on top. The task was formulated as a multi-label classification problem with 22 output units, including a no-tool class (i.e. background) as described by Sahu et al. (2016b). The network was trained using an Adam optimizer with a learning rate of 0.001 for 25 epochs. Assuming that tool usage transitions are smooth, linear temporal smoothing (Sahu et al., 2017) with a window of five frames is applied during inference in order to reduce false positives by suppressing stand alone detections.

4.8. RToolNet

RToolNet is a fine-tuned 50-layer residual network. After pre-training on ImageNet, the first 31 convolutional layers were frozen and only the remainder of the network was fine-tuned on the CATARACTS dataset using a decaying learning rate schedule. Furthermore, the approach makes heavy use of data augmentation to alleviate the strong correlation that is natural between video frames. The network was trained using a stochastic gradient descent with an initial learning rate of 0.05 and a momentum of 0.9. In the second submission, a weighted loss function was introduced which places more emphasis on training examples from underrepresented classes. This improved results slightly but also made the training more sensitive to inherent randomness, such as the choice of initial weights or training example order. We assume this to be the reason for the strong performance decrease observed for one tool between both submissions and note that this problem could be mitigated using an ensemble of networks trained with different random seeds.

4.9. CDenseNet

CDenseNet is based on DenseNet-169, and the last fully connected layer consists of 21 units for predicting the probability of the corresponding tool usage. To overcome the imbalance of the dataset, besides extracting 6 frames per second, more images were extracted for the rare tools, and a weighted binary softmargin loss function was adopted after converting all '0' labels in ground truth to '-1'. By this way, better performance was obtained for the rare tools, such as biomarker and Vannas scissors. To train the network, a stochastic gradient descent was used with a decreasing learning rate, initialized to 0.05, and a momentum of 0.9. Unlike other solutions, the CNN was not pre-trained on ImageNet: all weights were initialized randomly following a Gaussian distribution. Efficient DenseNet implementation (Pleiss et al., 2017) in PyTorch was used for accelerating the training procedure and improving the parameter utilization.

4.10. TUMCTNet

In the TUMCTNet solution, Inception-v4 was suitably modified and fine-tuned by introducing independent sigmoids as predictors for tool usage and by increasing the input size to 640 × 360 pixels to maintain the aspect ratio of the surgical video. To handle imbalance within multi-label settings, the co-occurrence of tools was considered for selecting the samples used for training: the label-set stratification proposed by Sahu et al. (2017) was used, which resulted in 46 label-sets. In addition to balancing the data-set, such an approach also exploits the relationship between tools during the surgery. During the training of the network, data-augmentation including limited random rotation (±10°), horizontal flipping, random scaling and center-cropping was used. Training relied on on a stochastic gradient descent with a learning rate of 0.001 and a momentum of 0.9. To improve temporal consistency of the results, temporal weighted averaging is performed during inference. An ensemble of two independently trained models is also employed to improve predictions.

4.11. CatResNet

The CatResNet model uses the 152-layer ResNet architecture for multi-label frame classification. The network was initialized with weights pre-trained on the ImageNet dataset and was further fine-tuned using the CATARACTS training videos (22 videos for training and 3 for validation). The videos were subsampled at 3 frames per second and half of the frames that do not feature any tool were discarded to match the frequency of the most common tool class, although the classes were not balanced further. The output of the network is a fully connected layer with 21 nodes with sigmoid activations and it was initialized with a Gaussian distribution with mean 0 and standard deviation 0.01 to be trained from scratch. During training, the input frames were re-shaped to 224 × 224 pixels and a random horizontal flip and random rotation within 25 degrees with mirror padding was performed to augment the data. The network was trained using stochastic gradient descent with a mini-batch of 8, a learning rate of 0.0001 and a momentum of 0.9 for a total of 10,000 iterations. For the first submission of this model, the predictions rely on the current frame alone and do not incorporate information from any other previous or following frame. A second submission was made which incorporates temporal smoothing as a post-processing step on the CNN predictions using a centered moving average kernel of size 5, however it does not achieve significantly better results.

4.12. TROLIS

The TROLIS solution differs from the competitors in two major aspects: (i) a classical computer vision algorithm is used to detect the biomarker (the rarest tool), and (ii) separate neural networks are trained for the rare tools and the rest. The training set was pruned first: the frames with video artifacts (tearing) were discarded, each 3 frames were averaged, and pixelwise similar frames were discarded. The tool categories were split into two: six rare tools and the remaining (regular) tools. For the regular tool identification, the average output of two Resnet-50 networks on frames resized to 256 × 256 pixels and one Resnet-50 network on frames resized to 512 × 512 pixels was used. These networks were optimized using stochastic gradient descents. For the rare tools, a new dataset was created: it consists of 3,000 (respectively 2,500) frames with (respectively without) rare tool labels. In addition to these frames from the training set, 1,200 frames from the test set, obtained by performing a forward pass using the three Resnet-50 networks, were used as negative samples. One of the networks was finetuned on this new dataset, and its output is used for rare tool identification. For the rarest tool (biomarker) detection, a classical computer vision algorithm is applied: it works by finding black blobs (tip of the marker) and white blobs (bulk of the marker) in each frame. It is assumed that the Mendez ring only appears in videos where the biomarker is present. Similarly, it is assumed that the needle holder only appears in videos with suture needle. Moreover, the first and last 0.5% frames of every test video is clipped. Finally, predictions are time averaged with a window of 45 frames.

4.13. CUMV

The CUMV solution relies on an ensemble of two CNNs with weights pre-trained on ImageNet: ResNet-101 and DenseNet-169. Each network takes as input a single frame from the surgical video, resized to 224 × 224 pixels, and outputs label predictions for the current frame. Both networks are trained independently with a stochastic gradient descent, using the cross-entropy loss. The learning rate was set to 0.001 for 6,000 iterations and then to 0.0001 for 5,000 iterations. During inference, a gate function (Hu et al., 2017) is used to combine the results of these two networks, which calculates the inner product of the normalized prediction confidences for each kind of tool.

4.14. DResSys

DResSys, developed at D-Wave, uses an ensemble of deep CNN networks to make predictions on individual video frames and then smooths these predictions across frames with a Markov random field. To extract video frames for training of the CNN ensemble, all frames within videos containing the rare tools (e.g. biomarker, Vannas scissors) were used, but in parts of the video with the most common tools, frames were sampled at a rate of only 6 frames/sec. Further, 40,000 frames were randomly selected at uniform rate from amongst training frames that have no tools. This process provided a total of ∼100,000 training images.

Frame-level predictors: In the first two submissions a single 50-layer Residual Network was trained and in subsequent submissions Inception-v4 and NASNet-A (Zoph et al., 2018) were trained in addition to ResNet. All parameters were initialized from pre-trained ImageNet models. Images of 540×960 pixels are used for ResNet-50 and Inception-v4, but since NASNet-A is a much larger network requiring much greater GPU memory, 270×480 images are used for this model. The final submission also uses one additional NASNet-A architecture with a larger image size of 337 × 600 pixels at input. The training data was augmented by randomly horizontally flipping and cropping images. All networks were trained with the Adam optimizer using a sigmoid cross-entropy loss except for the 337×600-pixel NASNet-A model that used a weighted sigmoid cross entropy loss. Training ran for at most 13 epochs with a batch size of 4. The learning rate for each network was chosen using cross validation. The prediction probability of each trained frame-level CNN is aggregated using a weighted geometric mean in which the weights were set using a grid search over the validation set.

Temporal smoothing: Several smoothing approaches were explored to capture the dependence of tool labels across consecutive frames. The first submissions were based on a simple median filtering method and the last submission includes a Markov random field (MRF) model. The MRF model provides a probability distribution across the time-dependent label space. Assume that $\boldsymbol{y} = \{y_1, y_2, \ldots, y_T\}$ represents the binary label vector for a given tool where $y_t = 1/0$ indicates the presence/absence of the tool in the t^{th} frame. The proposed MRF model has a chain-like structure and defines a conditional probability distribution $p(\boldsymbol{y}|\boldsymbol{x}) \sim \exp(-E(\boldsymbol{y};\boldsymbol{x}))$ for the label vector $\boldsymbol{y}$ given the video $\boldsymbol{x}$ using an energy function $E(\boldsymbol{y};\boldsymbol{x})$ given by

$$E(\boldsymbol{y};\boldsymbol{x}) = \sum_{t=1}^{T} a(s_t)y_t + \frac{w}{2}\sum_{t=1}^{T}\sum_{n\in N(t)} y_t y_n \quad , \qquad (1)$$

where $N(t) = \{t-19, t-17, \dots, t+19\}$ represents the set of neighboring nodes for the t^{th} frame, and provides long-range temporal connectivity. In Eq. (1), $a(s_t)$ is the bias for the t^{th} frame's label which is computed by shifting and scaling the output of the ensemble frame-level prediction score s_t at frame t. The scalar coupling parameter w in Eq. (1) enforces label agreement between neighboring frames. The w parameter and the shift and scale parameters of the linear map $a(s_t)$ were all set by a grid search and are shared for all the 21 tool categories. The MRF model, $p(\boldsymbol{y}|\boldsymbol{x})$, represents the joint probability distribution for all the labels in the temporal domain for a tool. Given this model, the marginal distribution $p(y_t = 1|\boldsymbol{x})$ is computed using a mean-field approximation (Jordan et al., 1999) and the resultant marginal probability is used as the prediction score for the t^{th} frame. Lastly, in order to process videos efficiently, the MRF model is formed in smaller segments of length ~20,000 frames.

4.15. LaTIM (organizers)

The LaTIM solution relies on an ensemble of CNNs, whose outputs are processed by an ensemble of RNNs. Convolutional and recurrent networks are trained sequentially using a novel boosting technique (Al Hajj et al., 2018). In a first submission, the CNN ensemble consists of one Inception-v4, one Inception-ResNet-v2 and one o_O network (Quellec et al., 2017b); the RNN ensemble consists of one LSTM (Hochreiter and Schmidhuber, 1997) and one GRU (Cho et al., 2014) network. In a second submission, a single CNN is used: NASNet-A. A different ensemble of RNNs, consisting of three LSTMs, is obtained. All networks are trained using the root mean square propagation algorithm. One major difference between both submissions is that RNNs are bidirectional in the first submission and unidirectional in the second, thus allowing online video analysis. Another difference is that a median filter is applied to each prediction signal in the second submission, for short-term temporal smoothing, whereas the RNNs are only used for long-term temporal analysis by design.

5. Results

A total of 27 submissions from 14 teams was received during the challenge period. Additionally, the organizers (LaTIM) submitted two solutions. A timeline of all these submissions is reported in Fig. 3. In order to establish a team ranking, the solution with maximal average AUC from each team was retained. Note that two solutions were evaluated after the challenge period, in virtue of the one week waiting rule: they were not used to establish the team ranking (see section 3.5). The leaderboard is reported in Table 7, together with the average AUCs and the detailed per-tool AUCs published on the CATARACTS website. This table also reports 95% confidence intervals (CIs) on the average AUCs, which were computed as follows: 1) CIs on the per-tool AUCs were computed using DeLong's method (DeLong et al., 1988), 2) their radii were then combined using the root mean square, assuming independence between tools. Each CI was used for a single comparison: is the corresponding solution significantly better than the following solution in the ranking? Results of this test are also reported in Table 7.

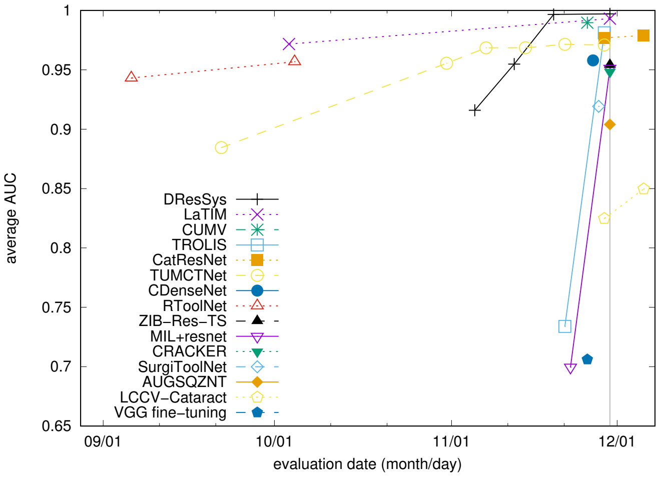


Figure 3: Timeline of solution evaluation — the gray vertical line indicates the challenge closing date. Evaluation dates and submission dates sometimes differed in virtue of the one week waiting rule.

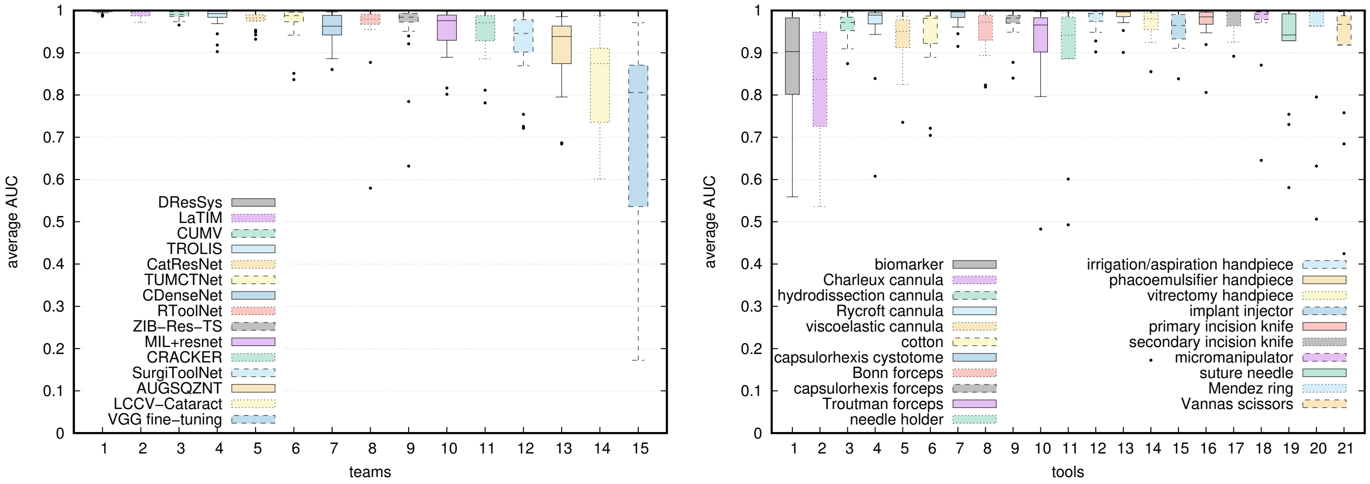
Per-tool AUCs are summarized in Fig. 4 using boxplots. Figure 4 (a) summarizes the performance of each solution: it appears that some solutions can detect all tools equally well while others fail for a few tools in particular. Figure 4 (b) summarizes how well each of these tools is detected by competing solutions: it appears that the Charleux cannula, the biomarker, the suture needle, the needle holder and the viscoelastic cannula are particularly challenging. On the contrary, the phacoemulsifier handpiece and the capsulorhexis cystotome are detected well by all solutions. ROC curves for simple and challenging tools are reported in Fig. 5.

For a deeper understanding of how each of these solutions analyze surgery videos, typical examples of temporal prediction signals are given in Fig. 6. One can easily notice which solutions include temporal smoothing techniques as post-processing steps (see Table 5). Another observation we can make is that the occurrence of false alarms is highly correlated in these signals: this is particularly clear in Fig. 6 (b).

Given the very good classification performance achieved by the top-raking solutions, we wondered whether or not they achieved human-level performance. To answer this question, we evaluated the competing solutions against the annotations of one expert only, before adjudication (see Fig. 7). We observed that the other human grader is always better than all competing solutions, in the sense that his sensitivity/specificity pair is above all ROC curves. A single exception was observed: for cotton usage detection, the DResSys algorithm is slightly better

team	training data selection	validation set
DResSys	6 frames per second	3 videos
LaTIM	30 frames per second	2 videos
CUMV	6 frames per second	5 videos
TROLIS	*frequent tools (3 CNNs):* torn frame removal, adaptive frame selection based on pixel differences *rare tools (5 CNNs):* 4200 negative frames (including 1200 test frames), 2500 positive frames	3 videos
CatResNet	3 frames per second	3 videos
TUMCTNet	0.8 frames per seconds	3 videos
CDenseNet	5 frames per second for frequent tools, 10 frames per second for rare tools	1/3 frames
RToolNet	5 frames per second, after removing 60% of frames without tools	5 videos
ZIB-Res-TS	6 frames per second, with labelset-based sampling (Sahu et al., 2017)	4 videos
MIL+resnet	15 frames per second	1/5 frames
CRACKER	15 frames per second	1/5 frames
SurgiToolNet	15 frames per second	2 videos
AUGSQZNT	10 frames per second	5 videos + selected frames with rare tools in 3 videos
LCCV-Cataract	24 frames per second	1/5 frames
VGG fine-tuning	15 frames per second	5 videos

Table 2: Training data and validation selection in the competing solutions.



(a) per-team boxplot

(b) per-tool boxplot

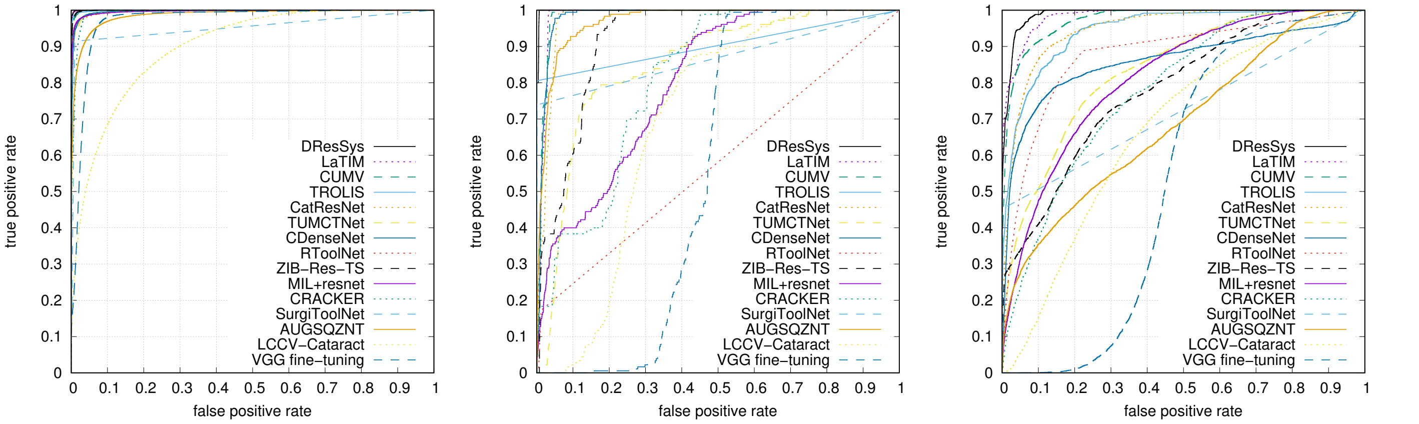
Figure 4: Boxplots of AUC scores grouped per team or per tool. Each box is drawn around the region between the first and third quartiles, with a horizontal line at the median value. Whiskers extend from the ends of each box to the most distant value which lies within 1.5 times the interquartile range. Black discs indicate outliers.

than the first human grader (see Fig. 7 (c)). To evaluate the cost of using automatic annotations rather than manual annotations, we computed the relative specificity decrease at equal sensitivity: results are reported in Table 8.

6. Discussion and Conclusions

We have presented the results of CATARACTS, the challenge on automatic tool annotation for cataract surgery. Given the high number of participants (14), we believe this challenge was a success. It is a unique opportunity to learn lessons that will guide the design of efficient surgery monitoring tools in the near future.

First, lessons can be learnt from the challenges noted by participants. All of them pointed out that the distribution of tools is highly unequal (see Fig. 2) and that tools in the same category are often visually similar to one another (cannulae, forceps, etc.). These problems motivated the use of data resampling strategies, to deal with class imbalance, and the design of adequate cost functions. It was also noted that video tearing artifacts appear at regular time intervals in videos. This problem motivated the use of time filtering techniques. Other properties of cataract surgery videos would probably have been listed as challenges in the pre-deep learning era: uneven illumination, zoom level variations, partial tool occlusion (only the tool tip is visible), and motion and out-of-focus blur. However, none of them were noted by participants: these problems are indeed handled well by CNNs coupled with adequate data augmentation strategies. On the other hand, other specificities of the CATARACTS dataset were exploited by participants to their advantage. First, tool usage generally does not change between consecutive frames. This factor also motivated the use



(a) phacoemulsifier handpiece

(b) biomarker

(c) Charleux cannula

Figure 5: Receiver-operating characteristic (ROC) curves. To save space, ROC curves are reported for three tools only: one frequent and well-detected tool (the phacoemulsifier handpiece) and two challenging tools (the biomarker and the Charleux cannula). Detecting the biomarker is challenging because there are few training samples. Detecting the Charleux cannula is challenging because this tool resembles the Rycroft cannula (in terms of shape and function).

team	random hor. flipping	random cropping	random scaling	random rotation	random shifting
DResSys	✓	✓			
LaTIM	✓		✓	✓	✓
CUMV	✓				
TROLIS	✓		✓	✓	✓
CatResNet	✓			✓	
TUMCTNet	✓	✓	✓	✓	
CDenseNet	✓	✓			
RToolNet	✓	✓		✓	
ZIB-Res-TS	✓		✓	✓	✓
MIL+resnet	✓		✓	✓	✓
CRACKER	✓		✓	✓	✓
SurgiToolNet	✓				
AUGSQZNT	✓	✓			
LCCV-Cataract					
VGG fine-tuning					

Table 3: Geometrical data augmentation in the competing solutions

of time filtering techniques. Second, tool usage usually follows precedence rules (e.g. phacoemulsification precedes implant injection) and the rarest tools are generally used in pairs to manage special events: bleeding (the suture needle and the needle holder), asymmetrical implant management (the biomarker and the Mendez ring), etc. These specificities motivated the use of (ad-hoc or general-purpose) temporal sequencers. However, the use of these temporal sequencers was to be used with caution, due to one specific challenge: tools in the same category are sometimes interchangeable. In particular, all forceps may be used to hold the suture needle, not only the 'needle holder'. In fact, one of the team that used recurrent neural networks (TROLIS) noted a performance increase after removing it.

The above-mentioned properties of the dataset and of the task at hand guided the design of the proposed solutions. Overall, most teams took the following steps to train their solutions: 1) selecting training frames in training videos, 2) downsampling these frames, 3) performing data augmentation, 4) selecting one or several CNNs pre-trained on ImageNet, 5) fine-tuning these CNNs on the selected video frames, through the minimization of a multi-label cost function, 6) optionally training a multi-CNN aggregation function and 7) optionally training a temporal sequencer. Selecting training frames (i.e. ignoring available training samples) and yet performing data augmentation (i.e. generating new training samples) may seem counter-intuitive. However, in many solutions, the decision to discard training frames was motivated by the need to balance classes. As for the general inference strategy, it can be summarized as follows: 1) resizing each test frame, 2) optionally performing data augmentation, 3) processing the resized frame with each CNN, 4) optionally aggregating the CNN predictions and 5) optionally running a temporal filter and/or sequencer. In other words, most participants followed the state-of-the-art approach for multi-label video sequencing using deep learning. It should be noted that no team designed a problem-specific CNN: all solutions relied on CNNs from the literature, with modifications in the final layers only. Beyond these general points, several lessons can be learnt by analyzing the differences between solutions. First, the following factors seem to positively impact the team ranking:

1. keeping full videos aside for validation, as illustrated in Table 2,
2. using data augmentation techniques, as illustrated in Table 3,
3. using the latest generation of CNNs, in particular their deepest versions, as illustrated in Table 4,
4. using multiple CNNs and/or RNNs, as illustrated in Table 4,
5. using temporal smoothing techniques, as illustrated in Table 5 and Fig. 6.

team	SqueezeNet (Iandola et al., 2016)	VGG-16 (Simonyan and Zisserman, 2015)	Inception-v3 (Szegedy et al., 2016)	Inception-v4 (Szegedy et al., 2017)	ResNet-34 (He et al., 2016)	ResNet-50 (He et al., 2016)	ResNet-101 (He et al., 2016)	ResNet-152 (He et al., 2016)	DenseNet-161 (Huang et al., 2017)	DenseNet-169 (Huang et al., 2017)	NASNet-A (Zoph et al., 2018)	image size	pre-training
DResSys				1		1					2	540×960 (× 2), 270×480, 337×600	ImageNet
LaTIM											1	331×331	ImageNet
CUMV							1			1		224×224	ImageNet
TROLIS						4						256×256 (× 3), 512×512	ImageNet
CatResNet								1				224×224	ImageNet
TUMCTNet				3								640×360 (× 3)	ImageNet
CDenseNet										1		540×960	no
RToolNet						1						540×960	ImageNet
ZIB-Res-TS						1						480×270	ImageNet
MIL+resnet			1		1							256×256 (early training stages: 128×128)	ImageNet
CRACKER					1							128×128	ImageNet
SurgiToolNet									1			224×224	ImageNet
AUGSQZNT	1											360×640	ImageNet
LCCV-Cataract			1									299×299	ImageNet
VGG fine-tuning		1										288×288	ImageNet

Table 4: Convolutional neural networks used in the competing solutions

team	test data augmentation	temporal smoothing
DResSys		Markov random field
LaTIM		LSTM (× 3), median filter
CUMV		
TROLIS		average filter
CatResNet		
TUMCTNet	center cropping	weighted average filter
CDenseNet		average filter
RToolNet		
ZIB-Res-TS		linear smoothing (Sahu et al., 2017)
MIL+resnet		rolling trimmed mean
CRACKER	4 versions of frame	median filter, zeroing of impossible predictions
SurgiToolNet		
AUGSQZNT	5 crops of frame	
LCCV-Cataract		
VGG fine-tuning		

Table 5: Post-processing techniques in the competing solutions

team	resampling	weighted loss	boosting	rare tool detector	co-occurrence analysis
DResSys		✓			
LaTIM			✓		✓
CUMV					
TROLIS				✓	✓
CatResNet					
TUMCTNet		✓			✓
CDenseNet	✓	✓			
RToolNet		✓			
ZIB-Res-TS	✓	✓			✓
MIL+resnet	✓				
CRACKER					✓
SurgiToolNet					
AUGSQZNT	✓				
LCCV-Cataract		✓			
VGG fine-tuning		✓			

Table 6: Strategies for class imbalance in the competing solutions

In fact, the winning team (DResSys) combined these five factors. The third lesson seems particularly important: solutions based on the recent NASNet-A architecture achieved top-ranking performance. On the other hand, the following factors do not seem to influence the team ranking: the number of selected training frames (see Table 2), the type of data augmentation (random cropping versus random affine transformations — see Table 3), the CNN's input image size (the CNN's default input size versus a larger size — see Table 4) or the use of test-time data augmentation (see Table 5). We observed that most methodological variations investigated by a single team were unsuccessful. Modeling the tool annotation task as a multi-class classification problem (LCCV-Cataract), rather than a multi-label one, was inefficient when more than two tools are used at the same time, which occurs frequently (see Fig. 2). Thresholding predictions as a post-processing step (SurgiToolNet), although important for use in production, decreased the solution's merit, evaluated by the area under the ROC curve (see Fig. 5

team	DResSys*	*LaTIM*	CUMV*	TROLIS*	CatResNet	TUMCTNet*	CDenseNet	RToolNet	ZIB-Res-TS	MIL+resnet*	CRACKER	SurgiToolNet	AUGSQZNT	LCCV-Cataract	VGG fine-tuning
rank	1		2	3	4	5	6	7	8	9	10	11	12	13	14
biomarker	0.9988	0.9847	0.9857	0.9026	0.9752	0.8511	0.9825	0.5797	0.9212	0.8018	0.8114	0.8690	0.9701	0.6983	0.5590
Charleux cannula	0.9892	0.9836	0.9735	0.9448	0.9490	0.8366	0.8603	0.8771	0.7846	0.8166	0.7814	0.7257	0.6867	0.6724	0.5359
hydrodissection cannula	0.9959	0.9873	0.9847	0.9840	0.9811	0.9570	0.9754	0.9717	0.9842	0.9704	0.9679	0.9091	0.9422	0.8743	0.9524
Rycroft cannula	0.9980	0.9946	0.9951	0.9924	0.9810	0.9907	0.9891	0.9908	0.9956	0.9791	0.9682	0.9709	0.9432	0.8391	0.6078
viscoelastic cannula	0.9865	0.9822	0.9776	0.9822	0.9423	0.9732	0.9349	0.9545	0.9506	0.9253	0.9120	0.9533	0.8248	0.7353	0.8588
cotton	0.9999	0.9986	0.9890	0.9842	0.9816	0.9854	0.9503	0.9759	0.9821	0.9702	0.9869	0.7213	0.9220	0.8893	0.7044
capsulorhexis cystotome	0.9999	0.9998	0.9987	0.9989	0.9976	0.9968	0.9966	0.9976	0.9933	0.9953	0.9911	0.9450	0.9832	0.9151	0.9609
Bonn forceps	0.9972	0.9949	0.9893	0.9942	0.9852	0.9825	0.9454	0.9726	0.9794	0.9574	0.9529	0.8934	0.9300	0.8188	0.8234
capsulorhexis forceps	0.9993	0.9981	0.9890	0.9845	0.9821	0.9879	0.9700	0.9888	0.9869	0.9759	0.9761	0.9779	0.9486	0.8774	0.8399
Troutman forceps	0.9898	0.9974	0.9917	0.9689	0.9752	0.9803	0.9237	0.9656	0.9827	0.9108	0.9020	0.9017	0.8744	0.7963	0.4826
needle holder	0.9945	0.9936	0.9846	0.9839	0.9500	0.9415	0.8859	0.9709	0.9395	0.8893	0.8853	0.9909	0.8990	0.6011	0.4929
irrigation/aspiration handpiece	0.9988	0.9989	0.9977	0.9976	0.9950	0.9947	0.9926	0.9913	0.9968	0.9925	0.9915	0.9279	0.9745	0.9019	0.9486
phacoemulsifier handpiece	0.9998	0.9998	0.9990	0.9993	0.9966	0.9969	0.9963	0.9971	0.9994	0.9966	0.9927	0.9526	0.9854	0.9006	0.9712
vitrectomy handpiece	0.9993	0.9719	0.9943	0.9960	0.9852	0.9924	0.9550	0.9932	0.9726	0.9778	0.9804	0.9958	0.8552	0.9244	0.1725
implant injector	0.9984	0.9939	0.9906	0.9935	0.9828	0.9852	0.9326	0.9644	0.9739	0.9486	0.9590	0.9172	0.9354	0.9108	0.8384
primary incision knife	0.9999	0.9965	0.9972	0.9933	0.9858	0.9961	0.9779	0.9848	0.9939	0.9801	0.9824	0.9674	0.9471	0.9195	0.8060
secondary incision knife	0.9997	0.9994	0.9995	0.9984	0.9984	0.9983	0.9911	0.9978	0.9995	0.9936	0.9889	0.9458	0.9632	0.9251	0.8917
micromanipulator	0.9989	0.9978	0.9940	0.9980	0.9897	0.9967	0.9886	0.9912	0.9917	0.9784	0.9710	0.9923	0.9815	0.6452	0.8706
suture needle	0.9987	0.9990	0.9861	0.9915	0.9320	0.9920	0.9420	0.9796	0.9920	0.9295	0.9284	0.7543	0.9383	0.7301	0.5810
Mendez ring	1.0000	0.9980	0.9999	0.9994	0.9966	0.9959	0.9629	0.9814	0.6317	0.9979	0.9986	0.9999	0.7952	0.9886	0.5064
Vannas scissors	0.9972	0.9842	0.9657	0.9182	0.9533	0.9705	0.9625	0.9673	0.9855	0.9893	0.9876	0.9925	0.6841	0.7579	0.4246
score (average AUC)	**0.9971**	**0.9931**	**0.9897**	**0.9812**	**0.9769**	**0.9715**	**0.9579**	**0.9568**	**0.9541**	**0.9513**	**0.9484**	**0.9192**	**0.9040**	**0.8248**	**0.7061**
lower bound of CI	0.9962	0.9923	0.9871	0.9737	0.9739	0.9653	0.9515	0.9481	0.9489	0.9433	0.9419	0.9004	0.8938	0.8123	0.6953
upper bound of CI	0.9981	0.9938	0.9916	0.9887	0.9799	0.9777	0.9643	0.9656	0.9592	0.9592	0.9549	0.9381	0.9142	0.8374	0.7169
better than the next ranked?	yes	yes	yes	no	yes	yes	no	no	no	no	yes	no	yes	yes	n/a

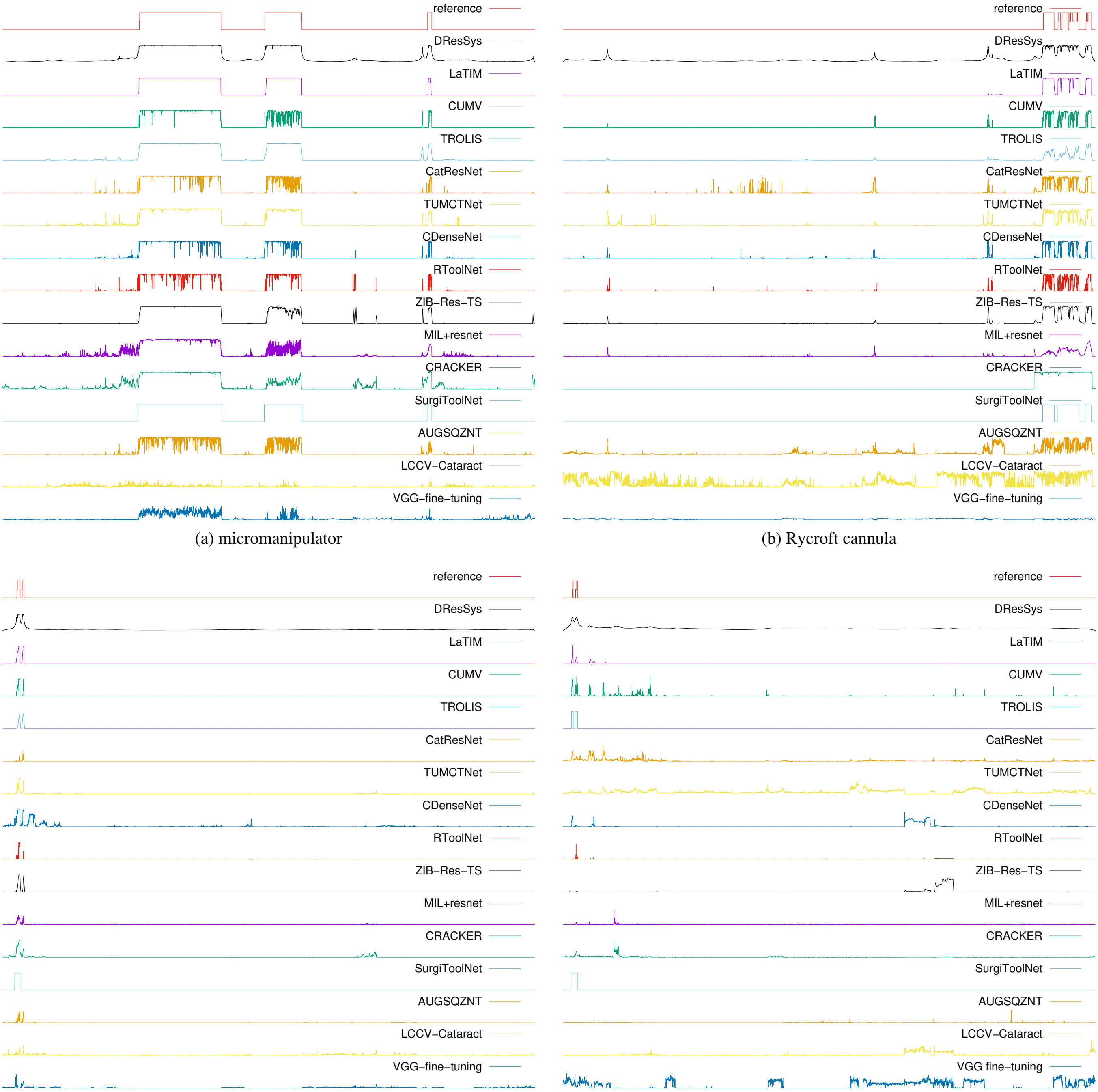
Table 7: Areas under the ROC curve (AUCs) for the retained solution of each team. To compare consecutive solutions in the ranking, 95% confidence intervals (CIs) on the average AUCs are included. CNN ensemble methods are indicated by an asterisk.

reference	expert 1	expert 2
DResSys	2.93 ± 0.84	1.91 ± 0.72
LaTIM	8.37 ± 2.33	5.58 ± 2.05
CUMV	13.52 ± 2.91	7.53 ± 2.18
TROLIS	19.02 ± 3.84	7.10 ± 2.09
CatResNet	24.74 ± 3.71	13.24 ± 2.81
TUMCTNet	26.15 ± 5.36	16.24 ± 5.32
CDenseNet	41.06 ± 5.55	22.78 ± 5.41
RToolNet	43.61 ± 7.02	26.39 ± 6.66
ZIB-Res-TS	27.97 ± 5.16	18.55 ± 5.08
MIL+resnet	41.36 ± 5.44	24.94 ± 5.25
CRACKER	34.31 ± 4.63	21.88 ± 4.45
SurgiToolNet	67.95 ± 6.95	40.59 ± 9.16
AUGSQZNT	66.13 ± 5.83	42.25 ± 7.61
LCCV-Cataract	68.91 ± 5.38	50.86 ± 5.44
VGG fine-tuning	70.00 ± 3.36	59.51 ± 4.08

Table 8: Relative specificity decrease, compared to the expert, at the same sensitivity. The relative specificity decrease is computed for all 21 tools and the average (± the standard error) is reported.

and 7). The use of a very simple classifier for rare but distinct tools like the biomarker (TROLIS) led to a very specific classifier (see Fig. 5 (b) and 6 (d)). However, like in the previous example, the use of binary predictions negatively impacted the score. Finally, we note that the most sophisticated solutions (MIL+resnet for instance) did not necessarily rank high, unless the general training procedure and the five success rules mentioned above were followed (like DResSys).

Compared to most medical image analysis challenges, one of CATARACTS' novelties was to offer participants the ability to submit multiple solutions over a long period of time (8 months). About half of the teams took advantage of this possibility during the last three months of that period (see Fig. 3). Several types of improvements were evaluated: improving data augmentation (tested by TUMCTNet between submissions 1 and 2 — noted "TUMCTNet 1 → 2"), selecting training images differently (DResSys 1 → 2, TROLIS 1 → 2, TUMCTNet 4 → 5 and MIL+resnet 1 → 2), replacing one CNN with another (LaTIM 1 → 2 and TUMCTNet 1 → 2), adding one or several CNNs (DResSys 2 → 3 & 3 → 4, TROLIS 1 → 2 and TUMCTNet 2 → 3 & 3 → 4 & 5 → 6), changing the input size of CNNs (TUMCTNet 4 → 5), redefining training images (DResSys 1 → 2, TROLIS 1 → 2, TUMCTNet 4 → 5 and MIL+resnet 1 → 2), redefining the loss function (DResSys 3 → 4, TUMCTNet 3 → 4, RToolNet 1 → 2, LCCV-Cataract 1 → 2), adding a temporal sequencer (DResSys 1 → 2, CatResNet 1 → 2, TUMCTNet 2 → 3 and MIL+resnet 1 → 2) and replacing this temporal sequencer with another (DResSys 2 → 3). The timeline in Fig. 3 reveals that consecutive submissions almost always led to a performance increase; the only exception was the last submission from the TUMCTNet team, although the decrease was minor. Increasing performance over time can be explained by the fact that participants progressively increased the complexity of their solution. It also indicates that



(a) micromanipulator

(b) Rycroft cannula

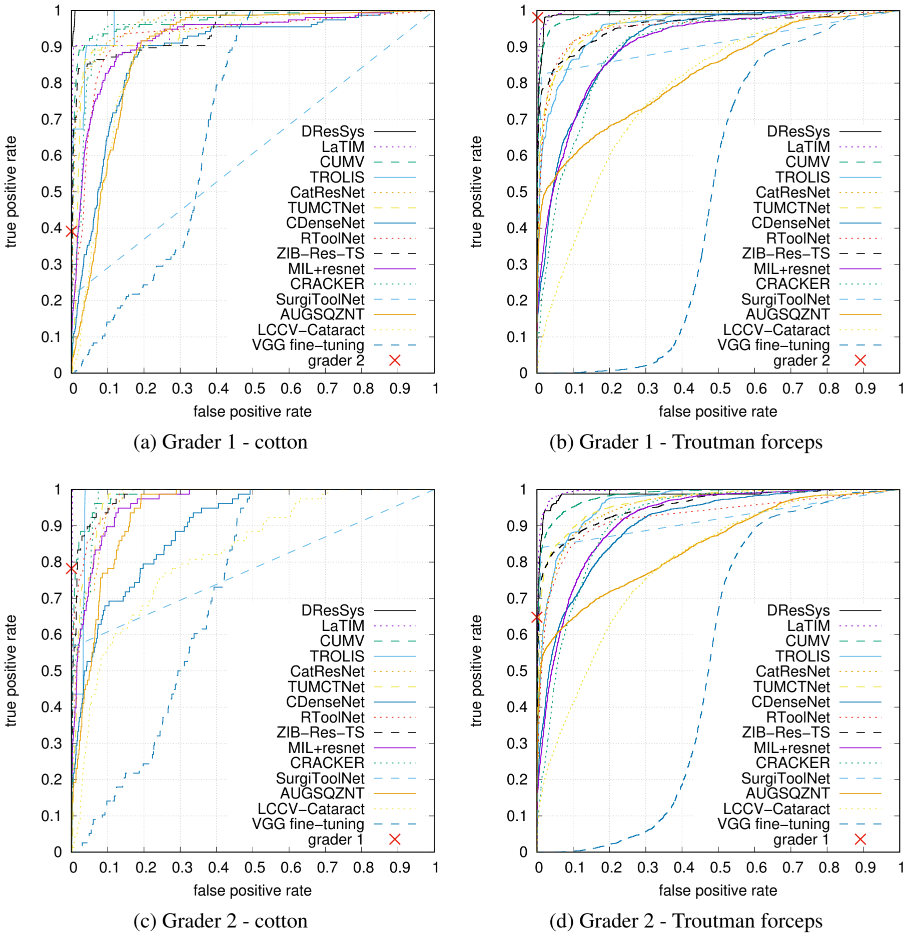
(c) secondary incision knife

(d) biomarker

Figure 6: Typical examples of temporal prediction signals. Predictions for the micromanipulator, the Rycroft cannula and the secondary incision knife are from a typical surgery (test video 6). Predictions for the biomarker are from a more complex surgery (test video 13).

participants progressively gained experience manipulating the training set and reading other teams' reports. On the down side, allowing multiple submissions introduced one unforeseen training bias: a few teams redefined their validation subset after detailed performance scores in the test set (per-tool AUC) revealed that some of the surgical tools did not appear in their training subset. On one hand, it helped correcting a careless mistake that could have been avoided by frequency counting in the training set. On the other hand, it can be regarded as training on the test set. These submissions were accepted anyway as they also included methodological novelties. As noted by other challenge organizers, challenge design is a delicate task (Maier-Hein et al., 2018).

This benchmarking study has one major limitation: solutions were only compared in terms of classification performance, while other aspects are also important. For instance, the ability



(a) Grader 1 - cotton

(b) Grader 1 - Troutman forceps

(c) Grader 2 - cotton

(d) Grader 2 - Troutman forceps

Figure 7: Receiver-operating characteristic (ROC) curves using the annotations of a single human grader, before adjudication, as reference standard. To save space, ROC curves are reported for two tools only. The sensitivity/specificity pair of the other expert is indicated by a red cross.

to analyze tool usage in real-time is of particular interest for the design of intraoperative decision support tools. Some participants (the AUGSQZNT team in particular) decided to design a lightweight solution that would run in real-time with limited hardware, which explains in part a lower ranking compared to those whose did not have that goal in mind. Given the setup of the challenge, it was not possible to compare computation times under identical conditions, so we did not analyze computational aspects in depth. A few lessons can be learnt anyway. First, computation times reported by most participants indicate that their solution can process several frames per seconds using one GPU, which would be enough in many applications. Second, it should be noted that most solutions allow online video analysis, in the sense that they don't need future information for inference. Of course, solutions relying on a symmetrical time filter (see Table 5) would infer predictions with a delay equal to the filter radius. However, this delay is usually less than a second, which would also be acceptable in many applications. Another aspect that would need further analysis is the independence on the acquisition hardware: to assess the generality of the proposed solutions, it would be useful to evaluate them on new datasets acquired with different microscopes, different cameras and/or different recorders.

As a final remark, we note that the classification performance of the proposed solutions is lower than that of a human expert (see Fig. 7). However, the performance of top-ranking solutions is very close (see Table 8). Given the limited performance decrease, an automated solution would clearly be a better option, especially in the context of intraoperative decision support: assuming a human interpreter can annotate tool usage in real time, he or she would have to dedicate one hundred percent of his or her time to that task, which would be prohibitive in the long term. Besides, we expect the performance of automated solutions to improve further should contextual information be available. In particular, additional video streams recording the surgical tray or the operating room in general could be considered. In conclusion, the CATARACTS challenge has demonstrated that the task of automated tool annotation in cataract surgery videos has virtually been solved, which paves the way for the introduction of innovative decision support technologies in the operating room, with benefits for both surgeons and patients.

7. Author contributions

Gwenolé Quellec and Mathieu Lamard initiated and designed the challenge. Hassan Al Hajj and Béatrice Cochener collected the dataset. Gwenolé Quellec and Hassan Al Hajj designed and implemented the statistical analysis. Béatrice Cochener guided the project from a clinical perspective. Gwenolé Quellec wrote the manuscript with substantial input from Pierre-Henry Conze, Hassan Al Hajj and Mathieu Lamard. All other authors designed solutions for the challenge and wrote a description of their solution in section 4; they also provided feedback for the rest of the manuscript.

8. Acknowledgments

Teresa Araújo is funded by the FCT grant contract SFRH/BD/122365/2016. Guilherme Aresta is funded by the FCT grant contract SFRH/BD/120435/2016. Adrian Galdran is with ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme, and National Funds through the FCT - Fundação para a Ciência e a Tecnologia within project CMUP-ERI/TIC/0028/2014. Aurélio Campilho is with project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016", financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF). Manish Sahu, Sabrina Dill, Anirban Mukhopadhyay and Stefan Zachow are funded by German Federal Ministry of Education and Research (BMBF) under the Project BIOPASS (grant Nr. 165V7257).

References

Agustinos, A., Voros, S., Oct. 2015. 2D/3D real-time tracking of surgical instruments based on endoscopic image processing. In: Proc Works CARE. Munich, Germany, pp. 90–100.

Al Hajj, H., Lamard, M., Charriere, K., Cochener, B., Quellec, G., Jul. 2017a. Surgical tool detection in cataract surgery videos through multi-image fusion inside a convolutional neural network. In: Proc IEEE EMBC. Seogwipo, South Korea, pp. 2002–2005.

Al Hajj, H., Lamard, M., Cochener, B., Quellec, G., Jul. 2017b. Smart data augmentation for surgical tool detection on the surgical tray. In: Proc IEEE EMBC. Seogwipo, South Korea, pp. 4407–4410.

Al Hajj, H., Lamard, M., Conze, P.-H., Cochener, B., Quellec, G., Jul. 2018. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. Med Image Anal 47, 203–218.

Allan, M., Ourselin, S., Hawkes, D., Kelly, J., Stoyanov, D., May 2018. 3-D pose estimation of articulated instruments in robotic minimally invasive surgery. IEEE Trans Med Imaging 37 (5), 1204–1213.

Alsheakhali, M., Eslami, A., Navab, N., Apr. 2016a. Detection of articulated instruments in retinal microsurgery. In: Proc IEEE ISBI. Prague, Czech Republic, pp. 107–110.

Alsheakhali, M., Eslami, A., Roodaki, H., Navab, N., 2016b. CRF-based model for instrument detection and pose estimation in retinal microsurgery. Comput Math Methods Med 2016, 1067509.

Attia, M., Hossny, M., Nahavandi, S., Asadi, H., Oct. 2017. Surgical tool segmentation using a hybrid deep CNN-RNN auto encoder-decoder. In: Proc IEEE SMC. Banff, AB, Canada, pp. 3373–3378.

Bernal, J., Tajkbaksh, N., Sánchez, F. J., Matuszewski, B. J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Córdova, H., Sánchez-Montes, C., Gurudu, S. R., Fernández-Esparrach, G., Dray, X., Liang, J., Histace, A., Jun. 2017. Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge. IEEE Trans Med Imaging 36 (6), 1231–1249.

Bodenstedt, S., Görtler, J., Wagner, M., Kenngott, H., Müller-Stich, B., Dillmann, R., Speidel, S., 2016 Feb-Mar. Superpixel-based structure classification for laparoscopic surgery. In: Proc SPIE Med Imaging. San Diego, CA, USA, p. 978618.

Bodenstedt, S., Wagner, M., Katić, D., Mietkowski, P., Mayer, B., Kenngott, H., Müller-Stich, B., Dillmann, R., Speidel, S., Feb. 2017. Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis. Tech. Rep. arXiv:1702.03684 [cs], Karlsruhe Institute of Technology.

Bouget, D., Allan, M., Stoyanov, D., Jannin, P., 2017. Vision-based and markerless surgical tool detection and tracking: a review of the literature. Med Image Anal 35, 633–654.

Casals, A., Amat, J., Laporte, E., Apr. 1996. Automatic guidance of an assistant robot in laparoscopic surgery. In: Proc IEEE ICRA. Vol. 1. Minneapolis, MN, USA, pp. 895–900.

Chang, P.-L., Rolls, A., Praetere, H., Poorten, E., Riga, C., Bicknell, C., Stoyanov, D., Jan. 2016. Robust catheter and guidewire tracking using B-spline tube model and pixel-wise posteriors. IEEE Robot Autom Lett 1 (1), 303–308.

Charrière, K., Quellec, G., Lamard, M., Martiano, D., Cazuguel, G., Coatrieux, G., Cochener, B., Nov. 2017. Real-time analysis of cataract surgery videos using statistical models. Multimed Tools Appl 76 (21), 22473–22491.

Chen, Z., Zhao, Z., Cheng, X., Oct. 2017. Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context. In: Proc IEEE CAC. Jinan, China, pp. 2711–2714.

Chmarra, M. K., Grimbergen, C. A., Dankelman, J., 2007. Systems for tracking minimally invasive surgical instruments. Minim Invasive Ther Allied Technol 16 (6), 328–340.

Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y., Oct. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In: Proc SSST. Doha, Qatar, pp. 103–111, arXiv: 1409.1259.

Cochener, B., Boutillier, G., Lamard, M., Auberger-Zagnoli, C., Aug. 2018. A comparative evaluation of a new generation of diffractive trifocal and extended depth of focus intraocular lenses. J Refract Surg 34 (8), 507–514.

Czajkowska, J., Pyciński, B., Juszczyk, J., Pietka, E., Apr. 2018. Biopsy needle tracking technique in US images. Comput Med Imaging Graph 65, 93–101.

de Silva, S. R., Evans, J. R., Kirthi, V., Ziaei, M., Leyland, M., 2016. Multifocal versus monofocal intraocular lenses after cataract extraction. Cochrane Database Syst Rev (12).

DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L., Sep. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44 (3), 837–845.

Donahue, J., Hendricks, L., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T., Apr. 2017. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell 39 (4), 677–691.

Du, X., Allan, M., Dore, A., Ourselin, S., Hawkes, D., Kelly, J., Stoyanov, D., 2016. Combined 2D and 3D tracking of surgical instruments for minimally invasive and robotic-assisted surgery. Int J Comput Assist Radiol Surg 11 (6), 1109–1119.

Du, X., Kurmann, T., Chang, P.-L., Allan, M., Ourselin, S., Sznitman, R., Kelly, J., Stoyanov, D., May 2018. Articulated multi-instrument 2-D pose estimation using fully convolutional networks. IEEE Trans Med Imaging 37 (5), 1276–1287.

Erie, J. C., Jan. 2014. Rising cataract surgery rates: Demand and supply. Ophthalmol 121 (1), 2–4.

Frikha, R., Ejbali, R., Zaied, M., 2016 Nov-Dec. Handling occlusion in augmented reality surgical training based instrument tracking. In: Proc IEEE AICCSA. Agadir, Morocco.

Funke, I., Bodenstedt, S., Riediger, C., Weitz, J., Speidel, S., Feb. 2018. Generative adversarial networks for specular highlight removal in endoscopic images. In: Proc SPIE Med Imaging. Houston, TX, USA, p. 1057604.

Furtado, A., Cheng, I., Fung, E., Zheng, B., Basu, A., Aug. 2016. Low resolution tool tracking for microsurgical training in a simulated environment. In:

Proc IEEE EMBC. Orlando, FL, USA, pp. 3843–3846.

García-Peraza-Herrera, L., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S., Sep. 2017. ToolNet: Holistically-nested real-time segmentation of robotic surgical tools. In: Proc IEEE IROS. Vancouver, Canada, pp. 5717–5722.

García-Peraza-Herrera, L., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S., Oct. 2016. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: Proc Works CARE. Athens, Greece, pp. 84–95.

Gessert, N., Schlüter, M., Schlaefer, A., May 2018. A deep learning approach for pose estimation from volumetric OCT data. Med Image Anal 46, 162–179.

He, K., Zhang, X., Ren, S., Sun, J., Jun. 2016. Deep residual learning for image recognition. In: Proc IEEE CVPR. Las Vegas, NV, USA, pp. 770–778.

Hochreiter, S., Schmidhuber, J., Nov. 1997. Long short-term memory. Neural Comput 9 (8), 1735–1780.

Hu, X., Yu, L., Chen, H., Qin, J., Heng, P.-A., Sep. 2017. AGNet: Attention-guided network for surgical tool presence detection. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Lecture Notes in Computer Science. Springer, Cham, pp. 186–194.

Huang, G., Liu, Z., Maaten, L. v. d., Weinberger, K. Q., Jul. 2017. Densely connected convolutional networks. In: Proc IEEE CVPR. Honolulu, HI, USA, pp. 2261–2269.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., Keutzer, K., Feb. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5mb model size. Tech. Rep. arXiv:1602.07360 [cs].

Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., Fei-Fei, L., Mar. 2018. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: Proc IEEE WACV. Lake Tahoe, NV, USA, pp. 691–699.

Jin, Y., Dou, Q., Chen, H., Yu, L., Heng, P.-A., Oct. 2016. EndoRCN: recurrent convolutional networks for recognition of surgical workflow in cholecystectomy procedure video. Tech. rep., The Chinese University of Hong Kong.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., Saul, L. K., Nov 1999. An introduction to variational methods for graphical models. Mach Learn 37 (2), 183–233.

Kaplowitz, K., Yazdanie, M., Abazari, A., 2018 Mar-Apr. A review of teaching methods and outcomes of resident phacoemulsification. Surv Ophthalmol 63 (2), 257–267.

Keller, B., Draelos, M., Tang, G., Farsiu, S., Kuo, A., Hauser, K., Izatt, J., 2018. Real-time corneal segmentation and 3D needle tracking in intrasurgical OCT. Biomed Opt Express 9 (6), 2716–2732.

Kelman, C. D., Jul. 1967. Phaco-emulsification and aspiration. A new technique of cataract removal. A preliminary report. Am J Ophthalmol 64 (1), 23–35.

Kessel, L., Andresen, J., Erngaard, D., Flesner, P., Tendal, B., Hjortdal, J., 2016. Indication for cataract surgery. Do we have evidence of who will benefit from surgery? A systematic review and meta-analysis. Acta Ophthalmol 94 (1), 10–20.

Kingma, D., Ba, J., May 2015. Adam: a method for stochastic optimization. In: Proc ICLR. San Diego, CA, USA.

Ko, S.-Y., Kim, J., Kwon, D.-S., Lee, W.-J., Aug. 2005. Intelligent interaction between surgeon and laparoscopic assistant robot system. In: Proc IEEE ROMAN. Nashville, TN, USA, pp. 60–65.

Krupa, A., Gangloff, J., Doignon, C., de Mathelin, M. F., Morel, G., Leroy, J., Soler, L., Marescaux, J., Oct. 2003. Autonomous 3-D positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. IEEE Transactions on Robotics and Automation 19 (5), 842–853.

Krwawicz, T., 1961. Intracapsular extraction of intumescent cataract by application of low temperature. Br J Ophthalmol 45 (4), 279–283.

Kügler, D., Jastrzebski, M., Mukhopadhyay, A., Jun. 2018. Instrument pose estimation using registration for otobasis surgery. In: Proc Works BIR. Leiden, The Netherlands, pp. 105–114.

Kurmann, T., Marquez, N., Du, X., Fua, P., Stoyanov, D., Wolf, S., Sznitman, R., Sep. 2017. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In: Proc MICCAI. Quebec City, Canada, pp. 505–513.

Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J., Eslami, A., Tombari, F., Navab, N., Sep. 2017. Concurrent segmentation and localization for tracking of surgical instruments. In: Proc MICCAI. Quebec City, Canada, pp. 664–672.

Lee, S., Tateno, K., Fuerst, B., Tombari, F., Fotouhi, J., Osgood, G., Johnson, A., Navab, N., Oct. 2017a. Mixed reality support for orthopaedic surgery. In: Proc IEEE ISMAR. Nantes, France, pp. 204–205.

Lee, S. C., Fuerst, B., Tateno, K., Johnson, A., Fotouhi, J., Osgood, G., Tombari, F., Navab, N., Oct. 2017b. Multi-modal imaging, model-based tracking, and mixed reality visualisation for orthopaedic surgery. Healthc Technol Lett 4 (5), 168–173.

Leppänen, T., Vrzakova, H., Bednarik, R., Kanervisto, A., Elomaa, A.-P., Huotarinen, A., Bartczak, P., Fraunberg, M., Jaaskelainen, J., Jun. 2018. Augmenting microsurgical training: Microsurgical instrument detection using convolutional neural networks. In: Proc IEEE CBMS. Karlstad, Sweden, pp. 211–216.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., Sánchez, C. I., Dec. 2017. A survey on deep learning in medical image analysis. Med Image Anal 42 (Supplement C), 60–88.

Loshchilov, I., Hutter, F., Apr. 2017. SGDR: Stochastic gradient descent with warm restarts. In: Proc ICLR. Toulon, France.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., Maier, O., Maier-Hein, K., Menze, B. H., Müller, H., Neher, P. F., Niessen, W., Rajpoot, N., Sharp, G. C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A. A., van der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., Kopp-Schneider, A., Jun. 2018. Is the winner really the best? A critical analysis of common research practice in biomedical image analysis competitions. Tech. Rep. arXiv:1806.02051 [cs].

Marban, A., Srinivasan, V., Samek, W., Fernández, J., Casals, A., Oct. 2017. Estimating position & velocity in 3D space from monocular video sequences using a deep neural network. In: Proc ICCV Works. Venice, Italy, pp. 1460–1469.

Mishra, K., Sathish, R., Sheet, D., Jul. 2017. Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2233–2240.

Miyawaki, F., Tsunoi, T., Namiki, H., Yaginuma, T., Yoshimitsu, K., Hashimoto, D., Fukui, Y., May 2009. Development of automatic acquisition system of surgical-instrument informantion in endoscopic and laparoscopic surgey. In: Proc IEEE ICIEA. Xi'an, China, pp. 3058–3063.

Niemeijer, M., Ginneken, B. v., Cree, M. J., Mizutani, A., Quellec, G., Sanchez, C. I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., Wu, X., Cazuguel, G., You, J., Mayo, A., Li, Q., Hatanaka, Y., Cochener, B., Roux, C., Karray, F., Garcia, M., Fujita, H., Abramoff, M. D., Jan. 2010. Retinopathy Online Challenge: Automatic detection of microaneurysms in digital color fundus photographs. IEEE Transactions on Medical Imaging 29 (1), 185–195.

Olson, R. J., Jan. 2018. Cataract Surgery From 1918 to the Present and Future - Just Imagine! Am J Ophthalmol 185, 10–13.

Pleiss, G., Chen, D., Huang, G., Li, T., van der Maaten, L., Weinberger, K. Q., Jul. 2017. Memory-efficient implementation of DenseNets. Tech. Rep. arXiv:1707.06990 [cs].

Popovic, M., Campos-Möller, X., Schlenker, M. B., Ahmed, I. I. K., Oct. 2016. Efficacy and safety of femtosecond laser-assisted cataract surgery compared with manual cataract surgery: A meta-analysis of 14 567 Eyes. Ophthalmol 123 (10), 2113–2126.

Primus, M., Schoeffmann, K., Böszörmenyi, L., Jun. 2016. Temporal segmentation of laparoscopic videos into surgical phases. In: Proc Works CBMI. Bucharest, Romania.

Quellec, G., Cazuguel, G., Cochener, B., Lamard, M., 2017a. Multiple-instance learning for medical image and video analysis. IEEE Rev Biomed Eng 10, 213–234.

Quellec, G., Charrière, K., Boudi, Y., Cochener, B., Lamard, M., Jul. 2017b. Deep image mining for diabetic retinopathy screening. Med Image Anal 39, 178–193.

Raju, A., Wang, S., Huang, J., Oct. 2016. M2CAI surgical tool detection challenge report. Tech. rep., University of Texas at Arlington.

Rathinam, A., Lee, Y., Ling, D., Singh, R., Sep. 2017. A review of image processing leading to artificial intelligence methods to detect instruments in ultrasound guided minimally invasive surgical procedures. In: Proc IEEE ICPCSI. Chennai, India, pp. 3074–3079.

Riaz, Y., Mehta, J. S., Wormald, R., Evans, J. R., Foster, A., Ravilla, T., Snellingen, T., Oct. 2006. Surgical interventions for age-related cataract. Cochrane Database Syst Rev (4), CD001323.

Rieke, N., Tan, D., Amat, d. S. F., Tombari, F., Alsheakhali, M., Belagiannis, V., Eslami, A., Navab, N., Dec. 2016a. Real-time localization of articulated surgical instruments in retinal microsurgery. Med Image Anal 34, 82–100.

Rieke, N., Tan, D., Tombari, F., Vizcáıno, J., di, S. F., Eslami, A., Navab, N., Oct. 2016b. Real-time online adaption for robust instrument tracking and pose estimation. In: Proc MICCAI. Athens, Greece, pp. 422–430.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kenngott, H., Speidel, S., Kopp-Schneider, A., Maier-Hein, K., Maier-Hein, L., Jun. 2018. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int J Comput Assist Radiol Surg 13 (6), 925–933.

Ryu, J., Moon, Y., Choi, J., Kim, H., 2018. A Kalman-Filter-Based common algorithm approach for object detection in surgery scene to assist surgeon's situation awareness in robot-assisted laparoscopic surgery. Journal of Healthcare Engineering 2018, 8079713.

Sahu, M., Moerman, D., Mewes, P., Mountney, P., Rose, G., Jan. 2016a. Instrument state recognition and tracking for effective control of robotized laparoscopic systems. Int J Mech Eng Robot Res 5 (1), 33–38.

Sahu, M., Mukhopadhyay, A., Szengel, A., Zachow, S., Oct. 2016b. Tool and phase recognition using contextual CNN features. Tech. Rep. arXiv:1610.08854 [cs.CV], Zuse Institute Berlin.

Sahu, M., Mukhopadhyay, A., Szengel, A., Zachow, S., Jun. 2017. Addressing multi-label imbalance problem of surgical tool detection using CNN. Int J Comput Assist Radiol Surg 12 (6), 1013–1020.

Sarikaya, D., Corso, J., Guru, K., Jul. 2017. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE Trans Med Imaging 36 (7), 1542–1549.

Seward, H. C., May 1997. Folding intraocular lenses: Materials and methods. Br J Ophthalmol 81 (5), 340–341.

Simonyan, K., Zisserman, A., May 2015. Very deep convolutional networks for large-scale image recognition. In: Proc ICLR. San Diego, CA, USA.

Su, Y.-H., Huang, K., Hannaford, B., Mar. 2018. Real-time vision-based surgical tool segmentation with robot kinematics prior. In: Proc IEEE ISMR. Atlanta, GA, USA, pp. 1–6.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A., Feb. 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proc AAAI. San Francisco, CA, USA, pp. 4278–4284.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., Jun. 2016. Rethinking the inception architecture for computer vision. In: Proc IEEE CVPR. Las Vegas, NV, USA, pp. 2818–2826.

Twinanda, A. P., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., Oct. 2016. Single- and Multi-Task Architectures for Tool Presence Detection Challenge at M2cai 2016. Tech. Rep. arXiv:1610.08851 [cs], University of Strasbourg.

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., Jan. 2017. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36 (1), 86–97.

Wang, S., Raju, A., Huang, J., Apr. 2017. Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In: Proc IEEE ISBI. Melbourne, Australia, pp. 620–623.

Wang, W., Yan, W., Fotis, K., Prasad, N. M., Lansingh, V. C., Taylor, H. R., Finger, R. P., Facciolo, D., He, M., Nov. 2016. Cataract surgical rate and socioeconomics: A global study. Invest Ophthalmol Vis Sci 57 (14), 5872–5881.

Wesierski, D., Jezierska, A., 2017 Aug-Sep. Surgical tool tracking by on-line selection of structural correlation filters. In: Proc EUSIPCO. Kos, Greece, pp. 2334–2338.

Wesierski, D., Jezierska, A., May 2018. Instrument detection and pose estimation with rigid part mixtures model in video-assisted surgeries. Med Image Anal 46, 244–265.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A., Dec. 2015. Describing videos by exploiting temporal structure. In: Proc IEEE ICCV. Santiago, Chile, pp. 4507–4515.

Ye, M., Zhang, L., Giannarou, S., Yang, G.-Z., Oct. 2016. Real-time 3D tracking of articulated tools for robotic surgery. In: Proc MICCAI. Athens, Greece, pp. 386–394.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., Nov. 2014. How transferable are features in deep neural networks? arXiv:1411.1792 [cs].

Zhao, Z., Voros, S., Weng, Y., Chang, F., Li, R., Dec. 2017. Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method. Comput Assist Surg (Abingdon) 22 (sup1), 26–35.

Zhou, M., Roodaki, H., Eslami, A., Chen, G., Huang, K., Maier, M., Lohmann, C., Knoll, A., Nasseri, M., 2017. Needle segmentation in volumetric optical coherence tomography images for ophthalmic microsurgery. Appl Sci 7 (8), 748.

Zia, A., Castro, D., Essa, I., Oct. 2016. Fine-tuning deep architectures for surgical tool detection. Tech. rep., Georgia Institute of Technology.

Zoph, B., Vasudevan, V., Shlens, J., Le, Q. V., Jun. 2018. Learning transferable architectures for scalable image recognition. In: Proc IEEE CVPR. Salt Lake City, UT, USA.