



HAL
open science

Micro-surgical anastomose workflow recognition challenge report

Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, et al.

► To cite this version:

Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, et al.. Micro-surgical anastomose workflow recognition challenge report. *Computer Methods and Programs in Biomedicine*, 2021, 212, pp.106452. 10.1016/j.cmpb.2021.106452 . hal-03414465

HAL Id: hal-03414465

<https://hal.science/hal-03414465>

Submitted on 10 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Highlights

- Micro-surgical anastomosis data set containing video, kinematic and workflow annotation.
- Challenge of surgical workflow recognition at different granularity levels
- Comparison of multiple deep learning based recognition methods

Journal Pre-proof

MIcro-Surgical Anastomose Workflow recognition challenge report

Arnaud Hualmé^{a,*}, Duygu Sarikaya^b, Kévin Le Mut^a, Fabien Despinoy^a,
Yonghao Long^{c,d}, Qi Dou^{c,d}, Chin-Boon Chng^{e,f}, Wenjun Lin^{e,f}, Satoshi
Kondo^g, Laura Bravo-Sánchez^h, Pablo Arbeláez^h, Wolfgang Reiterⁱ, Manoru
Mitsuishi^j, Kanako Harada^j, Pierre Jannin^{a,*}

^aUniv Rennes,INSERM, LTSI - UMR 1099, F35000, Rennes, France

^bGazi University, Faculty of Engineering; Department of Computer Engineering, Ankara,
Turkey

^cDepartment of Computer Science & Engineering, The Chinese University of Hong Kong

^dT Stone Robotics Institute, The Chinese University of Hong Kong

^eNational University of Singapore(NUS), Singapore, Singapore.

^fSouthern University of Science and Technology (SUSTech), Shenzhen, China.

^gKonica Minolta, Inc

^hCenter for Research and Formation in Artificial Intelligence, Department of Biomedical
Engineering, Universidad de los Andes, Bogotá, Colombia

ⁱWintegral GmbH

^jDepartment of Mechanical Engineering, the University of Tokyo, Tokyo 113-8656, Japan

Abstract

Background and Objective: Automatic surgical workflow recognition is an essential step in developing context-aware computer-assisted surgical systems. Video recordings of surgeries are becoming widely accessible, as the operational field view is captured during laparoscopic surgeries. Head and ceiling mounted cameras are also increasingly being used to record videos in open surgeries. This makes videos a common choice in surgical workflow recognition. Additional modalities, such as kinematic data captured during robot-assisted surgeries, could also improve workflow recognition. This paper presents the design and results of the “Micro-Surgical Anastomose Workflow recognition on training sessions” (MISAW) challenge whose objective was to develop workflow recognition models based on kinematic data and/or videos.

Methods: The MISAW challenge provided a data set of 27 sequences

*Corresponding author

Email addresses: arnaud.hualme@univ-rennes1.fr (Arnaud Hualmé),
pierre.jannin@univ-rennes1.fr (Pierre Jannin)

of micro-surgical anastomosis on artificial blood vessels. This data set was composed of videos, kinematics, and workflow annotations. The latter described the sequences at three different granularity levels: phase, step, and activity. Four tasks were proposed to the participants: three of them were related to the recognition of surgical workflow at three different granularity levels, while the last one addressed the recognition of all granularity levels in the same model. We used the average application-dependent balanced accuracy (AD-Accuracy) as the evaluation metric. This takes unbalanced classes into account and it is more clinically relevant than a frame-by-frame score.

Results: Six teams participated in at least one task. All models employed deep learning models, such as convolutional neural networks (CNN), recurrent neural networks (RNN), or a combination of both. The best models achieved accuracy above 95%, 80%, 60%, and 75% respectively for recognition of phases, steps, activities, and multi-granularity. The RNN-based models outperformed the CNN-based ones as well as the dedicated modality models compared to the multi-granularity except for activity recognition.

Conclusion: For high levels of granularity, the best models had a recognition rate that may be sufficient for applications such as prediction of remaining surgical time. However, for activities, the recognition rate was still low for applications that can be employed clinically. The MISAW data set is publicly available at www.synapse.org/MISAW to encourage further research in surgical workflow recognition.

Keywords: Surgical Process Model, Workflow recognition, Multi-modality, OR of the future

1 1. Introduction

2 Computer-assisted surgical (CAS) systems should ideally make use of a
3 complete and explicit understanding of surgical procedures. To achieve this, a
4 surgical process model (SPM) can be used. A SPM is defined as a "simplified
5 pattern of a surgical process that reflects a predefined subset of interest of

6 the surgical process in a formal or semi-formal representation” [1]. The SPM
7 methodology is used for various applications, such as operating room optimization
8 and management [2, 3], learning and expertise assessment [4, 5], robotic assistance
9 [6], decision support [7], and quality supervision [8].

10 According to Lalys et al. [9], a surgical procedure can be decomposed on
11 several levels of granularity ,e.g., phases, steps, and activities. Phases are the
12 decomposition of a surgical procedure into the main periods of intervention (e.g.,
13 resection). Each phase is broken down into multiple steps corresponding to a
14 surgical objective (e.g., to resect the pouch of Douglas). A step is composed of
15 several activities that describe the physical actions (namely action verbs,e.g., cut)
16 performed on specific targets (e.g., the pouch of Douglas) by specific surgical
17 instruments (e.g., a scalpel). This initial definition was improved at a lower
18 granularity level to take into account information closed to kinematic data [10]:
19 surgemes and dexemes. A surgeme represents a surgical motion with explicit
20 semantic meaning (e.g., grab), and a dexeme is a numerical representation of
21 the sub-gestures necessary to perform a surgeme.

22 In early publications [2, 3, 4, 5, 6, 7, 8, 10], SPMs were manually acquired
23 by human observers. However, this solution has several drawbacks: It is costly
24 concerning human resources, time-consuming, observer-dependent, and errors
25 could be made. In [11], the authors noted that for the annotation of a peg
26 transfer task, the mean duration to manually annotate one minute of video was
27 around 13 minutes, and 65 annotation errors were counted for 60 annotations
28 although the task was less susceptible to subjective interpretation than a surgical
29 operation. To overcome these issues, [11] proposed an automatic annotation
30 method based on the information extracted from a virtual reality simulator.
31 Even though this is a promising solution to limit human annotation, it requires
32 information that could be complicated to obtain in surgical practice, such as
33 the interactions between the instruments and anatomical structures. Other
34 solutions are currently being studied to reduce the amount of manual annotation
35 as transfer learning from simulated data to real data [12] or from a limited
36 amount of annotated data [13].

37 Despite these innovative methods, automatic and online recognition of surgical
38 workflows is mandatory to bring context-awareness CAS applications inside the
39 operating room. Various machine learning and deep learning methods have been
40 proposed to recognize different granularity levels such as phases [3, 14, 15] ,
41 steps [16, 17] ,and activities [6, 18]. According to the type of surgery, different
42 modalities could be used for workflow recognition. For manual surgery, unless it
43 is possible to add multiple sensors, workflow recognition is generally restricted
44 to video-only modalities [3, 16, 18]. In the case of robot-assisted surgery (RAS),
45 kinematic information is easily available. It is expected that multi-modal data will
46 lead to easier automatic recognition methods, as is the case for the combination
47 of video and eye gaze information [17] or the combination of video and kinematic
48 information based on RAS data [19]. However, some methods based on RAS
49 data sets propose video-only methods [20, 21] or kinematic-only methods [10, 13].

50 The “Micro-Surgical Anastomose Workflow recognition on training sessions”
51 (MISAW) challenge provided a unique data set for online automatic recogni-
52 tion of multi-granularity surgical workflows using kinematic and stereoscopic
53 video information on a micro-anastomosis training task. The participants were
54 challenged to develop uni-granularity (with phases, steps, or activities) and/ or
55 multi-granularity workflow recognition models.

56 **2. Methods: Reporting of Challenge Design**

57 This section describes the challenge design through an explanation of the
58 organization, the mission, the data set, and the assessment method of the
59 challenge.

60 *2.1. Challenge organization*

61 The MISAW challenge was a one-time event organized as part of EndoVis
62 for MICCAI2020 online. It was organized by five people from three different
63 institutions: Arnaud Huaultmé, Kévin Le Mut, and Pierre Jannin from the
64 University of Rennes (France), Duygu Sarikaya from Gazi University (Turkey),

65 and Kanako Harada from the University of Tokyo (Japan). The challenge
66 was partially funded by the ImPACT Program of the Council for Science,
67 Technology and Innovation, Cabinet Office, Government of Japan. All challenge
68 information was made available to the participants through the Synapse platform:
69 www.synapse.org/MISAW .

70 Participation in the challenge was subject to the following policies: Partici-
71 pants had to submit a fully automatic method using kinematic and/or video data.
72 The data that could be used for the training were restricted to the data provided
73 by the organizers and publicly available data sets, including pre-trained networks.
74 The publicly available data sets only covered data that were available to everyone
75 when the MISAW data set was released. The results of all participating teams
76 were announced publicly on the challenge day. Challenge organizers and people
77 from the organizing institutions could participate but were not eligible for the
78 competition.

79 The participating teams had to provide the following elements: the method's
80 outputs, a write-up, and a Docker image allowing the organizers to verify the
81 outputs provided. Due to the COVID-19 crisis, a pre-recorded talk was also
82 mandatory to limit technical issues during the challenge day (online event). All
83 technical information (how to create a Docker image, the output format, etc.)
84 was provided to the participants during the challenge on the challenge platform.
85 The participants could submit multiple results and Docker images. However, only
86 the last submission was officially counted to compute the challenge results. No
87 leader-board or evaluation results were provided before the end of the challenge.

88 The challenge schedule was as follows: The training and the test data sets
89 were released on June 1st and August 24th 2020 respectively. Submissions were
90 accepted until September 23rd (23:59 PST). The results were announced October
91 4th during the online MICCAI2020. The complete data set was released with
92 this paper at: www.synapse.org/MISAW

93 The organizers' evaluation scripts were publicly available on the challenge
94 platform. Participating teams were encouraged (but not required) to provide
95 their code as open access.

96 *2.2. Mission of the challenge*

97 The objective of the challenge was to automatically recognize the workflow
98 of an anastomose performed during training sessions using video and kinematic
99 data. The challenge was composed of four different tasks according to the
100 granularity level recognized. Three of these tasks were uni-granularity surgical
101 workflow recognition, i.e., the model had to recognize one of the three available
102 granularity levels. They were noted task 1 for phase recognition, task 2 for step
103 recognition, and task 3 for activities recognition. The last task, noted task 4,
104 was a multi-granularity surgical workflow recognition, i.e., recognition of the
105 three granularities with the same model. Each task offers a different difficulty,
106 from the easiest (task 1) to the most complex (task 4). Due to the hierarchical
107 decomposition of SPM, phases are more distinct between them than activities.
108 For example, an activity could have the same action verb and surgical tool but
109 only differ by the target, whereas two phases represent two distinct main periods
110 of the operation. Task 4 brings together the difficulty of the three other ones.
111 However, due to the hierarchical structure of the workflow description, it could
112 help the recognition of the lower granularity. For example, some activities can
113 only be present in specific phases or steps.

114 The challenge data were provided by a robotic system used to realize micro-
115 surgical anastomosis on artificial blood vessels through a stereoscopic microscope.
116 Such micro-surgical anastomosis is performed in neurosurgery and plastic surgery.
117 The surgical robotic technologies developed for micro-surgical anastomosis can
118 be applied to other robotic surgeries requiring dexterous manipulation on small
119 targets. Automatic recognition of this task is an essential step to help the
120 realization of this task or to increase robotic autonomy from manual to shared
121 control or full automation [22].

122 The final biomedical application was robotic micro-surgical suturing of the
123 dura mater during endonasal brain tumor surgery. Both applications were similar
124 in the use of a robotic system, the microscopic dimension of the targets, and the
125 surgical gestures. Furthermore, the automatic workflow recognition associated
126 with analysis methods could help the surgical skill training by providing online

127 feedbacks. For example, on a subset of the MISAW dataset, Huauilmé et al. [4]
 128 have identified sequences of activities, called surgical signatures, specific to the
 129 expertise or to common mistakes.

130 2.3. Challenge data set

131 The challenge data set was composed of 27 sequences of micro-surgical
 132 anastomosis on artificial blood vessels performed by 3 surgeons and 3 engineering
 133 students. It was divided into a training data set composed of 17 cases and a test
 134 data set composed of 10 cases. The splitting of the data set was done to have a
 135 similar ratio of expertise in each data set (Tableau 1). A case was composed of
 136 kinematic data, a video, and workflow annotation. The latter was not provided
 137 to participants for the test cases.

Participant	Training cases				Test cases	
	Surgeon 2	Surgeon 3	Student 1	Student 2	Surgeon 1	Student 3
nb case	3	4	6	4	4	6

Table 1: Training and test case splitting

138 2.3.1. Data acquisition

139 The video and kinematic data were synchronously acquired at 30 Hz by a
 140 high-definition stereo-microscope (960x540 pixels) and a master-slave robotic
 141 platform [23], respectively, by the Department of Mechanical Engineering of the
 142 University of Tokyo. The kinematic data were recorded by encoders mounted
 143 on the two robotic arms. The kinematic data consisted of $x, y, z, \alpha, \beta, \gamma$. The
 144 homogeneous transformation matrices for each robotic instrument were calculated
 145 as in equations 1 and 2. The kinematic files also contained information about
 146 the grip and the output grip voltage.

$$H_{right} = T_x(x)T_y(y)T_z(z)R_x\left(\frac{1}{18}\pi\right)R_y(\alpha)R_x\left(\beta - \frac{5}{9}\pi\right)R_y(\gamma) \quad (1)$$

147

$$H_{left} = T_x(x)T_y(y)T_z(z)R_x\left(-\frac{1}{18}\pi\right)R_y(\alpha)R_x\left(\beta + \frac{1}{18}\pi\right)R_y(\gamma) \quad (2)$$

Phases	Steps	Activities			
		Verb	Target	Instrument	
Suturing	Needle holding	Catch	Needle	Needle holder	
Knot Tying	Suture making	Give slack	Wire		
	Suture handling	Hold	Both artificial vessel		
	1° knot	Insert	Left artificial vessel		
	2° knot	Loosen completely	Right artificial vessel		
	3° knot	Loosen partially	Long wire strand		
		Make a loop	Short wire strand		
		Pass through	Wire loop		
		Position	Knot		
		Pull			

Table 2: MISAW vocabulary.

148 The workflow annotation was acquired manually by two non-medical observers
 149 from the MediCis team of the LTSI Laboratory from the University of Rennes.
 150 The observers used the software "Surgery Workflow Toolbox [annotate]" provided
 151 by the IRT b<>com [24] to annotate the phases, steps, and activities (action
 152 verb, target, and instrument) of each robotic arm according to an annotation
 153 protocol. The vocabulary contained 2 phases, 6 steps, 10 action verbs, 9 targets,
 154 and 1 surgical instrument (Table 2). The protocol described how to recognize
 155 each phase, step, and activity of each robotic arm by giving a definition, start
 156 and end point, and graphical illustration. For example, the step "suture making"
 157 was defined by "insert and pull the needle into artificial vessels." The start point
 158 was the "beginning of the needle insertion on one vessel," the stop point was
 159 "the needle completely pass through both vessels." This is illustrated in Figure
 160 1. The complete annotation protocol is available in Supplementary Material C.

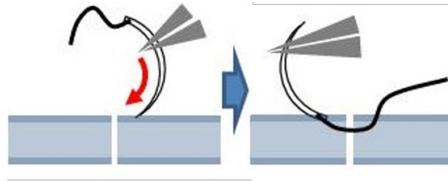


Figure 1: Representation of the beginning (left) and the end (right) of the "suture making" step.

161 Each case was annotated by both observers independently and harmonized
 162 by the following protocol (Figure 2). An automatic merging was performed when
 163 the transition difference between both observers was less than one second (b in
 164 Figure 2). Here, the transition between red and blue components was inferior to
 165 the threshold, so the automatic merging took the mean. The transition between
 166 the blue and the green components took longer than one second, so no decision
 167 was made. The merging sequence came back to each observer separately to
 168 refine uncertain transitions (c). A second automatic annotation was performed
 169 with a threshold of 0.5 seconds (d). Finally, all remaining uncertainties were
 170 harmonized by a consensus between both observers.

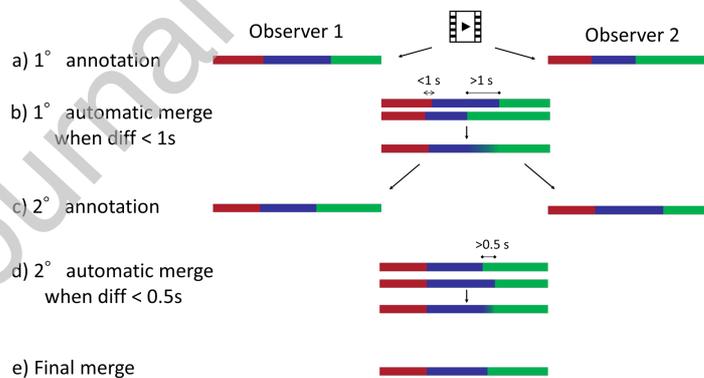


Figure 2: Harmonization protocol used to merge the annotations made by two observers.

171 2.3.2. Data pre-processing

172 We pre-processed the videos and initial workflow annotations to have consis-
 173 tent and synchronized data for each case. In the videos, the boundary between

174 the left image and the right image was not consistent (Figure 3, i.e., the position
 175 of the centerline was a little different within and between the trials. We removed
 176 40 pixels from the center of the stereoscopic image to have two images of 460x540
 177 pixels. The final video resolution was 920x540 pixels.

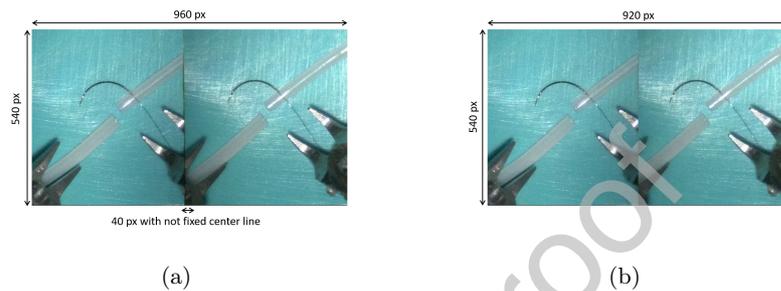


Figure 3: Comparison between the initial video (3a) and pre-processed video (3b)

178 The software “Surgery Workflow Toolbox [annotated]” produced a description
 179 of sequences where each element was characterized by the beginning and the
 180 end in milliseconds. We modified it to provide a discrete sequence synchronized
 181 at 30Hz with the kinematic data. When no phase, step, or activity occurred,
 182 the term "Idle" was added. For each timestamp, we provided the following in-
 183 formation: timestamp_number, phase_value, step_value, verb_Left_Hand, tar-
 184 get_Left_Hand, instrument_Left_Hand, verb_Right_Hand, target_Right_Hand,
 185 and instrument_Right_Hand.

186 2.3.3. Sources of errors

187 The main source of errors was the manual workflow annotation, which was
 188 observer-dependent. We limited these errors through the double annotations
 189 and the harmonization.

190 The second possible source of errors came from an acquisition issue. During
 191 acquisition, some timesteps were not acquired in either the video and kinematic
 192 information. This did not affect the synchronization of the data but could create
 193 activities not present in the procedural description. The impacted cases were

194 2-3, 4-2, 4-4, and 5-3.

195 Finally, due to some system problems during acquisition, the grip data were
 196 doubtful. If the system worked correctly, 0 meant "open" and -6 meant "close,"
 197 but maybe values were under -6 in some trials.

198 These sources of errors were communicated to the participants with the
 199 training data set. The participants did not report any other issues.

200 2.4. Assessment method

201 2.4.1. Metrics

202 To assess the methods proposed by participants, we used a balanced version
 203 of the application-dependent scores [25] of the classic metric used in the workflow
 204 recognition: accuracy, precision, recall, and F1.

205 Our data sets had a high class unbalance, for example, the phase "Idle"
 206 represented around 2% of the frames in both data sets (Figure 4). To give the
 207 same importance to each class, we decided to use balanced scores.

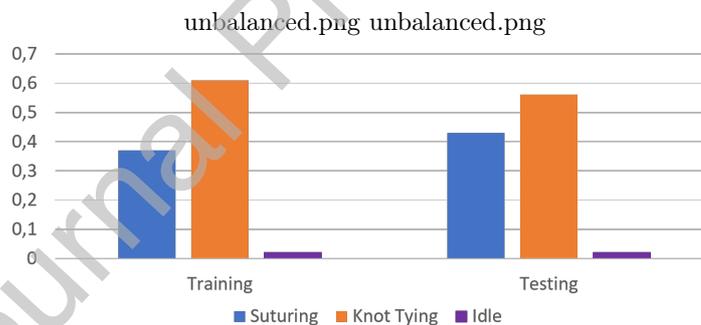


Figure 4: Phase distribution in training and testing data sets

208 Generally, frame-by-frame scores were used. This type of score assumes that
 209 the ground truth is frame perfect. However, this is not possible with manual
 210 annotation. Moreover, a clinical application does not need to be 100% accurate
 211 at a frame resolution. Application-dependent scores re-estimate classic scores
 212 using acceptable delay thresholds for a transitional window (Figure 5). When
 213 the transition on the predicted sequence occurs with a transition delay TD

214 inferior to an acceptable delay d centered into the real transition, all frames are
 215 considered correct. Here, this was the case for the transition between the blue
 216 and green components. If the transition was different (case between red and
 217 green components in the prediction sequence) or outside this transition delay,
 218 no modification was done. We fixed the acceptable delay d at 500 ms, which
 219 corresponded to half of the duration used for the first automatic merge (Figure
 220 2).

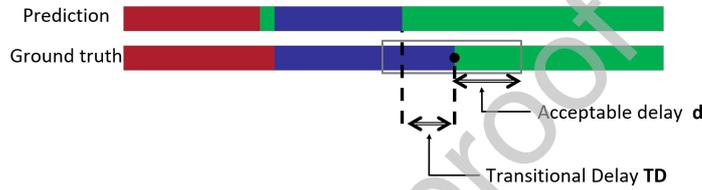


Figure 5: Definition of acceptable and transitional delay used to compute the average-dependent scores

221 2.4.2. Ranking method

222 We used a metric-based aggregation on the balanced application-dependent
 223 accuracy (AD-Accuracy) for the ranking. For each participant, we aggregated
 224 the metric values over all test cases and aggregated overall metrics to obtain a
 225 final score. We used a metric-based aggregation according to the conclusion of
 226 [26], who reported this type of aggregation as one of the most robust.

227 For the phase and step recognition (Tasks 1 and 2), the ranking score for
 228 algorithm a_i was computed as follows:

$$s_{uni}(a_i) = \frac{\sum_{t=0}^T \text{balance_accuracy_case_}t}{T} \quad (3)$$

229 Activity recognition (Task 3) consisted of recognizing the action verb, target,
 230 and instrument of each robotic arm. The ranking metric was computed as follows,

231 with each component (i.e., $s_{verb_Left}(a_i)$) computed with Equation 3:

$$s_{activity}(a_i) = \frac{1}{6} * (s_{verb_Left}(a_i) + s_{Target_Left}(a_i) + s_{Instrument_Left}(a_i) + s_{verb_Right}(a_i) + s_{Target_Right}(a_i) + s_{Instrument_Right}(a_i)) \quad (4)$$

232 For the multi-granularity recognition (Task 4) the ranking score was the
233 mean of each uni-granularity score:

$$s_{multi}(a_i) = \frac{s_{phase}(a_i) + s_{step}(a_i) + s_{activity}(a_i)}{3} \quad (5)$$

234 All multi-granularity recognition models were also ranked in each uni-granularity
235 task to highlight the differences between the models.

236 In the case of missing results, we considered results as good as a total random
237 recognition. For example, for 3-class problem, the missing result would be set to
238 $1/3$, and for a 12-class problem, it would set to $1/12$.

239 The ranking stability was assessed by testing different ranking methods. If,
240 the ranking was not stable according to the method chosen, a tie between the
241 different teams was pronounced.

242 The ranking computation and analysis were performed with the ChallengeR
243 package provided by [27].

244 3. Results: Reporting of the Challenge Outcomes

245 3.1. Challenge submission

246 At the end of September 2020, we counted 24 individual participants registered
247 to the MISAW challenge and 325 downloads of the 9 files available (the Synapse
248 platform did not give statistics by file). Five competing teams and one non-
249 competing team completed their submissions for the challenge.

250 3.2. Information on participating teams and corresponding methods

251 In this section, we will present information on each team, the methods
252 they used, and which tasks they participated in. The presentation is made in
253 alphabetical order of the competing teams and not in consideration of their
254 ranking. Table 3 provides a summary of participants methods.

255 *3.2.1. MedAIR*

256 The MedAIR team was composed of Yonghao Long and Qi Dou from the
257 Department of Computer Science and Engineering at the Chinese University of
258 Hong Kong. They participated in the phase and step recognition tasks.

259 The MedAIR team used both the video frames and the kinematic data of
260 the left and right robotic arms, treating them separately because different arms
261 may conduct different actions to jointly complete a task.

262 They extracted high-level features from video frames using an 18-layer residual
263 convolutional network [28] followed by a fully connected layer and a ReLU non-
264 linearity layer applied at the end, yielding a 128-dimension spatial feature vector.
265 To learn the temporal information of the video data, they adopted a temporal
266 convolution network (TCN) [29], i.e., an encoder-decoder module, to further
267 capture the information across frames, generating the representative spatial-
268 temporal visual features. For the kinematic data, they first normalized the
269 variables into $[-1, +1]$, and then they used the TCN and long short-term memory
270 (LSTM) [30] in parallel to learn and model the complex information of the left
271 and right arms separately, yielding spatial-temporal motion features.

272 After acquiring the encoded high-level features from the video stream and
273 kinematic data of the left and right arms, they used a graph convolutional network
274 to further learn the joint knowledge among the multi-modal data. Considering
275 that the visual information and left/right kinematic information contained fruitful
276 interactions and relationships, they designed a graph convolutional network
277 (GCN) with three node entities corresponding to the video, left kinematics, and
278 right kinematics, with all three nodes connected to each other. Initialized with
279 these three modalities, the node features of the GCN were updated by receiving
280 the message from neighbor connected node features and then encoding stronger
281 information in the newly generated node features. Then, the authors max pooled
282 the features from the three nodes and forwarded them into a fully connected
283 layer to get the prediction results of the workflow recognition. For more details,
284 you can refer to [31].

285 Two different approaches were employed to further enhance the temporal
286 consistency of the workflow recognition. The authors filtered out the frames with
287 low prediction probabilities using a median filter and leveraged the information
288 of the preceding 600 frames with 30 fps. They also employed an online post-
289 processing strategy (PKI) [32] that leveraged the workflow of the phases and
290 steps. For example, the steps followed a specific order: "Needle holding", "Suture
291 making", "Suture handling", "1 knot", "2 knot", and "3 knot", and it were not
292 likely to be reversed or shuffled.

293 3.2.2. NUSControl Lab

294 The NUSControl Lab team was composed of Chin Boon Chng¹, Wenjun
295 Lin^{1,2}, Jiaqi Zhang¹, Yaxin Hu¹, Yan Hu¹, Liu Jiang Jimmy², and Chee Kong
296 Chui¹. The participants noted with "1" were from the National University of
297 Singapore (NUS), Singapore, Singapore; the participants noted with "2" were
298 from the Southern University of Science and Technology, Shenzhen, China. This
299 team participated in the multi-granularity task. As described in the subsection
300 "Ranking method" (2.4.2), the model was also ranked in each uni-granularity
301 task.

302 The NUSControl Lab team used both the video and kinematic data. They
303 first extracted the features of the video frames using EfficientNet [33]. Then, they
304 employed an LSTM module to model the sequential dependencies of the video
305 data. The authors hypothesized that the kinematic data were specifically related
306 to the verbs and steps. With this motivation, they employed another LSTM
307 module to model the sequential features of the left and right arm kinematic data,
308 which was then concatenated and fed into a fully connected layer to predict the
309 verbs (left and right) and the steps. Their network model was based upon the
310 work of Jin et al. [34].

311 The authors also employed a post-processing step that made used of the
312 workflow observations to further improve the predictions. For example, if a knot
313 is to be tied, a loop must first be made, followed by pulling the wire. Thus, the
314 verb "make a loop" could be used to indicate when a new knot is being tied.

315 Similarly, the verb “pull” could be used to indicate when the new knot has been
316 completed. The authors proposed to mark the verb “make a loop” as a transition
317 signal to the next knot and “pull” as a completion signal of this knot. If the
318 model classified the current task to be “making a loop” and the phase turned to
319 knot tying, the knot step was incremented. This knot step was identified to start
320 from the previous “pull” prediction and continue until the next “pull” prediction.

321 3.2.3. SK

322 The SK team was composed of Satoshi Kondo from Konica Minolta, Inc. This
323 team participated in the multi-granularity task. As described in the subsection
324 "Ranking method" (2.4.2), the model was also ranked in each uni-granularity
325 task.

326 The SK team used the video data, kinematic data, and time information
327 as the input for the model. The video frame features were extracted using a
328 50-layer ResNet [28] pre-trained with the ImageNet data set, which led to a
329 2,048-dimension feature vector. While the team used only the left stereo video
330 frame, the kinematic data features: $x, y, z, \alpha, \beta, \gamma$, and grip collected from the
331 left and right arms were used, leaving the output voltage for the grip feature
332 out. The kinematic data were normalized with the mean and standard deviation
333 values for each dimension and then fed to two fully connected layers.

334 The team also employed the frame number as a means of time information.
335 The frame number was divided by 10,000. The feature vector of the input
336 image, the feature vector of the kinematic data, and the frame number were
337 concatenated, which led to a 2,063-dimension feature vector for a single frame.
338 Then, the author performed multi-granularity recognition wherein the network
339 learned the tasks, i.e., phase, step, and activity. For each activity, the verb, the
340 target, and the tool for the left and right arms were learned, which resulted in
341 a total of eight classes. The loss function was the summation of softmax cross-
342 entropy for these eight classes, and the team employed a Lookahead optimizer
343 [35].

344 3.2.4. *UniandesBCV*

345 The UniandesBCV team was composed of Laura Bravo-Sánchez, Paola Ruiz
346 Puentes, Natalia Valderrama, Isabela Hernández, Cristina González, and Pablo
347 Arbeláez. All members were from the Center for Research and Formation
348 in Artificial Intelligence and the Biomedical Computer Vision Group at the
349 Universidad de los Andes, Colombia. This team participated in all proposed
350 tasks.

351 The UniandesBCV team only used the video data and proposed a model that
352 leveraged the implicit hierarchical information in the surgical workflow. The
353 model presented by the authors was based on SlowFast [36], a neural network
354 that uses a slow and a fast pathway to model semantic and temporal information
355 within videos. To accomplish this discrimination of information, each of the
356 pathways analyzed the video at a different sampling rate. The slow pathway
357 used a low frame rate with a large number of channels, while the fast pathway
358 employed a high frame rate and only a fraction of the channels. To make a
359 prediction based on the complete information (semantic and temporal), the fast
360 pathway fused with the slow one using several lateral connections at different
361 points of the network.

362 The authors first extracted the features of the video frames using ResNet-50
363 backbone [28] and fed the features of 64-frame windows into a SlowFast model
364 adapted for multi-task training that was pre-trained on the Kinetics data set
365 [37]. The authors explored different multi-task hierarchical groupings: The
366 first model simultaneously predicted both phases and steps, the second model
367 predicted activities, and the last model predicted all the components of the
368 multi-granularity recognition. During training, the team also introduced an
369 extra term to the loss function for optimizing the task added at each grouping
370 and balanced the relevance of each task by associating each of the loss's terms
371 to a weight. The authors reported that merging all the components of the
372 multi-granularity recognition tasks improved the learning ability of the model
373 and obtained more accurate predictions.

374 3.2.5. *Wr0112358*

375 Team wr0112358 was composed of Wolfgang Reiter from Wintegral GmbH.
376 This team participated in all proposed tasks.

377 Team wr0112358 only used the video data, reporting that the kinematic data
378 did not significantly contribute to the performance of the model. The team
379 also explored different architectures, including ResNet50 [28] and multi-stream
380 Siamese networks with temporal pooling [38], but reported that due to the high
381 imbalance and the relatively small size of the data set, the complex architectures
382 resulted in overfitting. The author also ruled out using an LSTM approach for
383 the same reason.

384 The team decided to employ a multi-task convolutional neural network [15]
385 and extracted the features of the video frames using a DenseNet121 CNN with
386 data augmentation and regularization, which reduced overfitting. The author
387 enhanced this architecture with task-wise early stopping [39] and also reported
388 that using either of the stereo video frames resulted in a similar performance.

389 3.2.6. *IMPACT*

390 The IMPACT team was the non-competing team due to the presence of
391 challenge organizers on it. This team was composed of Arnaud Hualmé and
392 Fabien Despinoy both from Rennes University, INSERM, LTSI - UMR 1099 and
393 Duygu Sarikaya from Gazi University, Faculty of Engineering, Department of
394 Computer Engineering. The team participated in all proposed tasks.

395 The IMPACT team used both the right video frames and kinematic data
396 and proposed a multi-modal architecture. The authors applied a pre-processing
397 step to the input data. While the right video images were rescaled from 460x540
398 to 224x224 and the pixel values were normalized by subtracting the mean of
399 each channel over the training set and scale between [0,1], the authors applied
400 a z-normalization to center the kinematic data to 0 with a standard deviation
401 equal to 1. To make the training step faster, the data were downsampled to 5Hz.

402 Then, each input modality was processed into a dedicated network branch
403 to leverage the different types of data. While the video frames were passed to

404 a VGG19 network [40], the kinematic data were passed to an adapted ResNet
 405 network [41]. The last convolutional layer of each modality branch was finally
 406 concatenated into a common branch before being split again into separated
 407 workflow branches containing their own activation layers (1 for the phase and
 408 step recognition, 6 for the activity recognition, and 8 for the multi-granularity
 409 recognition).

410 The VGG19 network was initialized with the weights of a pre-trained model
 411 on the ImageNet data set. Since the MISAW data set was acquired in phantom
 412 surgical settings, the IMPACT team retrained only the last two layers to refine
 413 the network for this task. Regarding the kinematic branch, the network was
 414 trained "from scratch" without any previous weight configuration. In the end,
 415 the training was achieved using an Adam optimizer and a starting learning rate
 416 of 0.0001.

Team name	Tasks	Modalities	Networks	Post processing
MedAIR	1, 2	V + K	CNN, TCN, RNN	PKI
NUSControl Lab	4	V+K	CNN, RNN	Workflow
SK	4	V+K	CNN	
UniandesBCV	All	V	SlowFast, CNN	
Wr0112358	All	V	CNN	
IMPACT	All	V+K	CNN	

Table 3: Participating teams and their methods summary. On modalities, V is for video, K for Kinematic.

417 3.3. Workflow recognition results

418 Even if the participants submitted the method outputs for each test case, all
 419 of the following results were computed on organizer hardware via the provided
 420 Docker images. We did not detected any fraud attempts in the results provided
 421 by the participants.

422 This section only presents the results used for the ranking. Detailed results

423 by sequence and task of each participating team are available in Supplementary
 424 Material B.

425 3.3.1. Task 1: Phase recognition

426 Phase recognition is a three-class task. We received 4 uni-granularity and 5
 427 multi-granularity models for this task; the latter were identified with the addition
 428 of "_multi" at the end of the team name.

429 Figure 6 presents the results of all algorithms for each test sequence. The
 430 average AD-Accuracy by sequence was between 77.7% and 84.7%, which demon-
 431 strated that the recognition difficulty was similar between the sequences, except
 432 for sequence 5_6. However, we noticed that for all the test cases, 2 models had
 433 an AD-Accuracy lower than 65%.

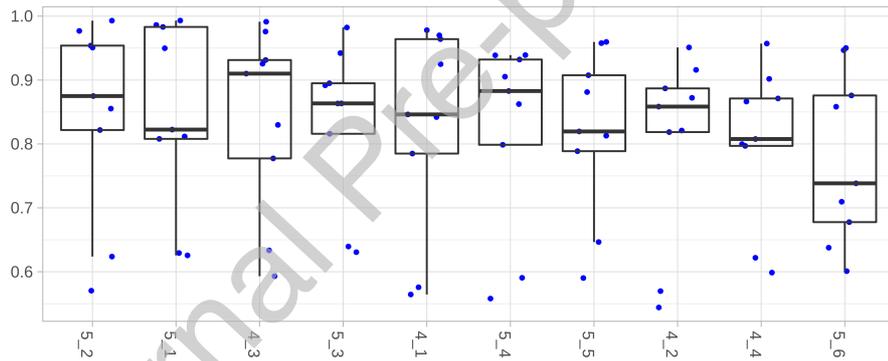


Figure 6: Phase recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

434 Figure 7 presents the average AD-Accuracy for each model. The MedAIR
 435 team got an average AD-Accuracy of 96.53%. The multi-granularity models of
 436 the UniandesBCV and SK teams presented results lower than 65%. Overall,
 437 only the uni-granularity model of IMPACT had an outlier lower than 70%, while
 438 the average AD-Accuracy was 80.66%. Is it also interesting to note that the
 439 multi-granularity model of this team was slightly better than the uni-granularity
 440 one.

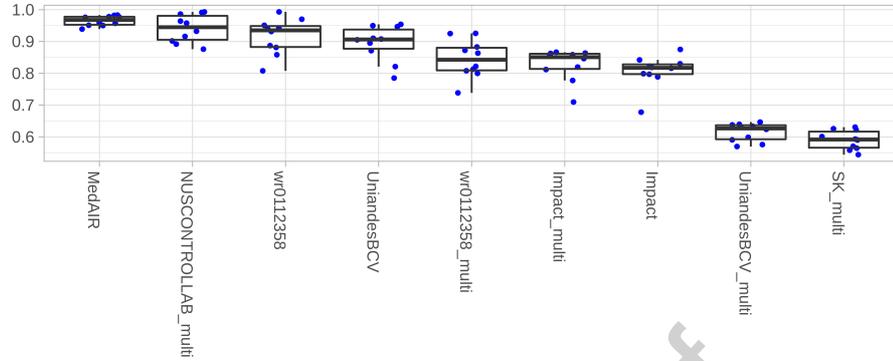


Figure 7: Average phase recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

441 Figure 8 presents the different rankings according to the method chosen.
 442 For the phase recognition, the choice of method did not influence the ranking,
 443 except for the multi-granularity models of teams IMPACT and wr0112358, which
 444 swapped the fifth and sixth places.

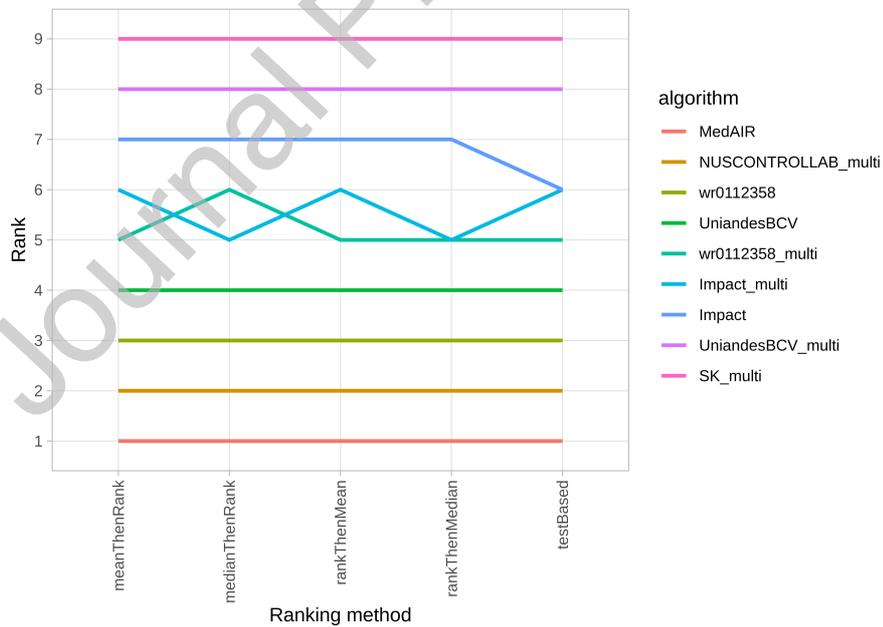


Figure 8: Phase recognition ranking stability through different ranking methods.

445 3.3.2. Task 2: Step recognition

446 Step recognition is a 7-class task. We received 4 uni-granularity and 5 multi-
 447 granularity models for this task; the latter were identified with the addition of
 448 "_multi" at the end of the team name.

449 The average AD-Accuracy by sequence was between 51.2% and 64.4% (Figure
 450 9). Contrary to the phase recognition, there was no sequence with a significantly
 451 lower score. We noticed that for all sequences, at least one model outperformed
 452 the others.

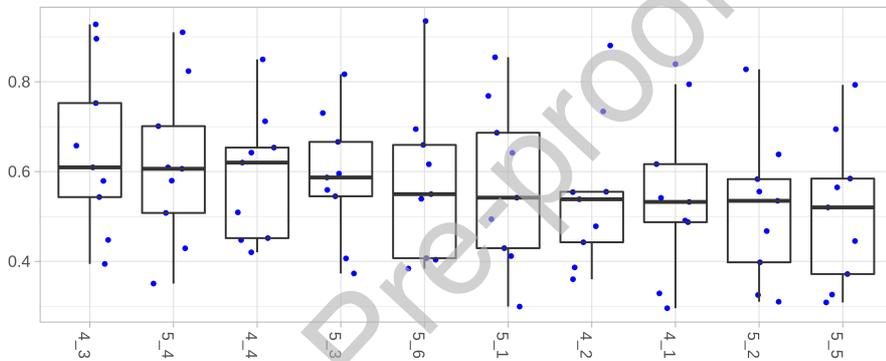


Figure 9: Step recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

453 In figure 10, we could identify this team as MedAIR, which obtained an
 454 average AD-Accuracy of 84.02%. Three models had results lower than 50%: the
 455 uni-granularity model of the IMPACT team and the multi-granularity models
 456 of the UniandesBCV and SK teams. Only the multi-granularity model of
 457 NUSControl Lab had disparate results according to the sequences.

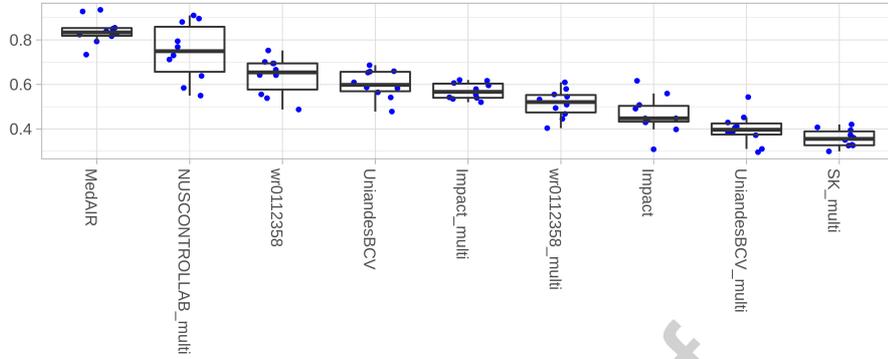


Figure 10: Average step recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

458

The ranking method chosen did not impact the final rank (Figure 11).

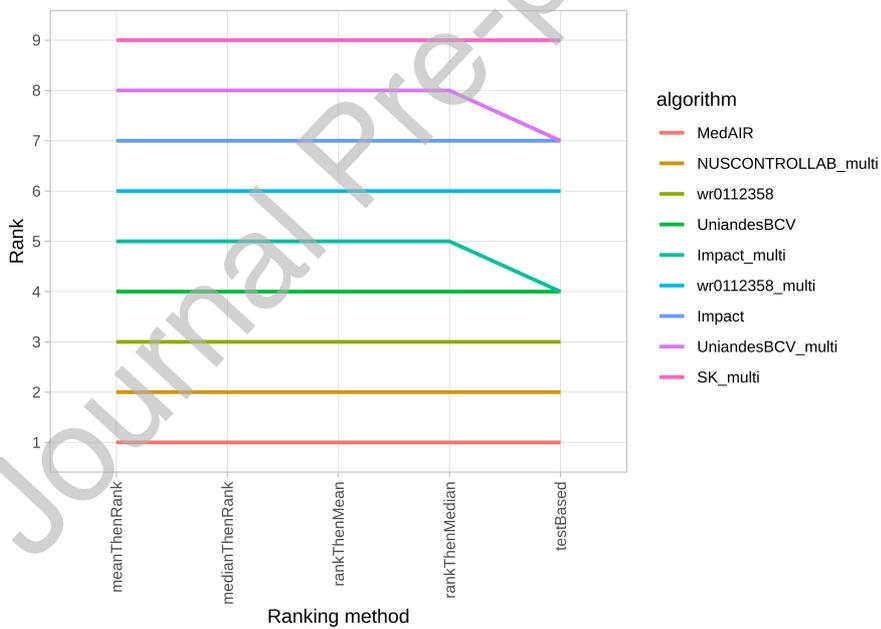


Figure 11: Step recognition ranking stability through different ranking methods.

459 *3.3.3. Task 3: Activity recognition*

460 The activity recognition consisted of recognizing 6 components, i.e., the verb,
 461 target, and instrument for the left and right arms. Each component was an
 462 11-, 10-, and 2-class problem, respectively. We received 3 uni-granularity and 5
 463 multi-granularity models for this task; the latter were identified with the addition
 464 of "_multi" at the end of the team name.

465 The average AD-Accuracy score by sequence was between 55.1% and 63.4%
 466 (Figure 12). As for the step recognition, all sequences had similar results.
 467 However, for session 4_4, one model had an AD-Accuracy lower than 40%.

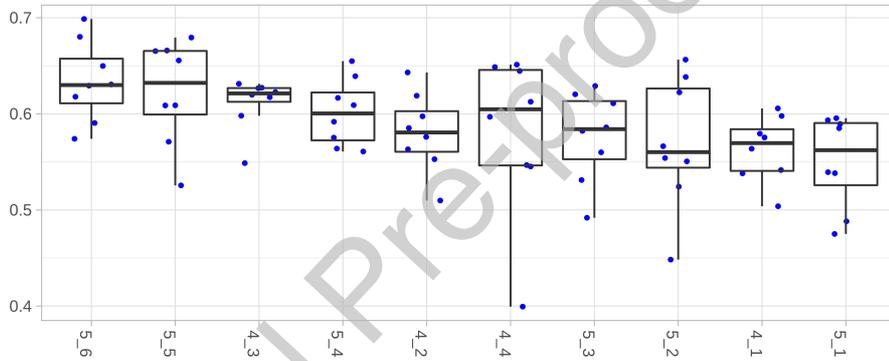


Figure 12: Activity recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

468 The average AD-Accuracy by model was between 52.4% and 61.6% 13. Four
 469 models, three of which were multi-granularity ones, had results over 60%.

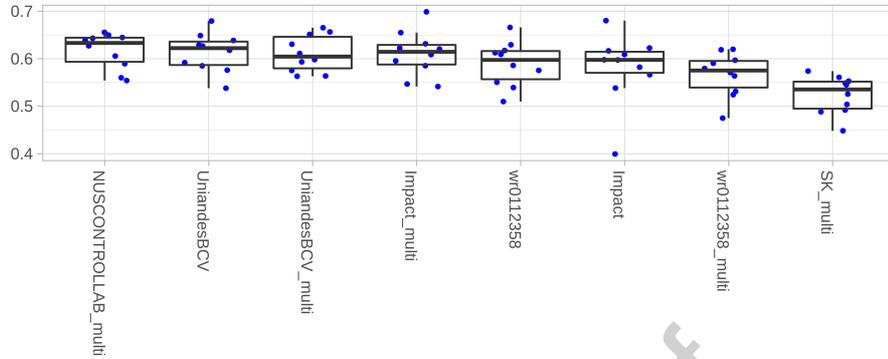


Figure 13: Average activity recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

470 According to the ranking method (Figure 14), the ranking was always different
 471 for the top four models. For this task, we defined a tie between the NUSControl
 472 Lab and UniandesBCV teams (IMPACT was a non-competitive team).

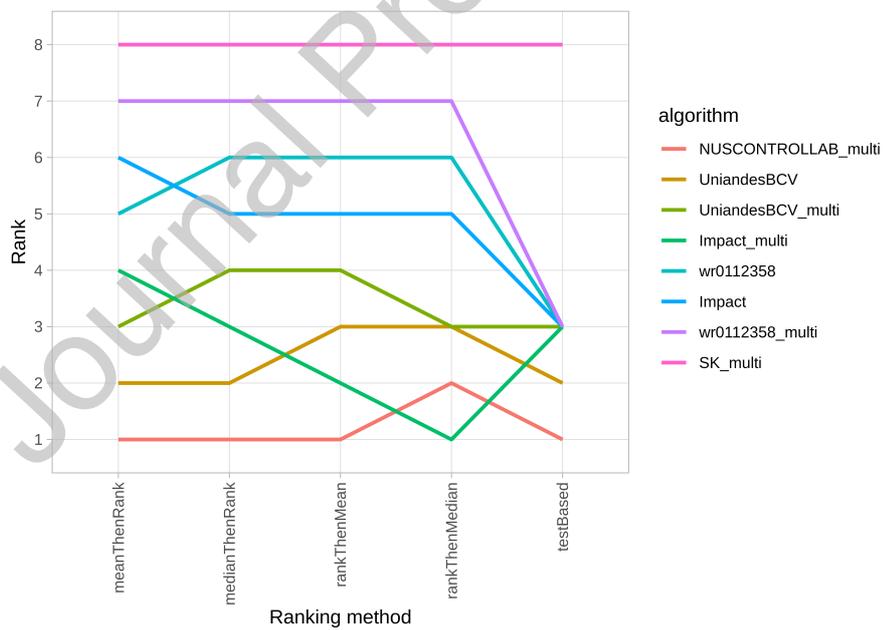


Figure 14: Phase recognition ranking stability through different ranking methods.

473 3.3.4. Task 4: Multi-granularity recognition

474 Task 4 consisted of recognizing the phase (a 3-class problem), the steps (a
 475 7-class problem), and the verb, target, and instrument for the left and right arms
 476 (an 11-, 10-, and 2- class problem, respectively) on a unique model . Of the 6
 477 teams, 5 proposed a model for this task.

478 The average AD-Accuracy score by sequences was between 59.6% and 66.4%
 479 (Figure 15). Surprisingly, these results were slightly better than those for the
 480 activity recognition although this task also demanded recognition of phases and
 481 steps.

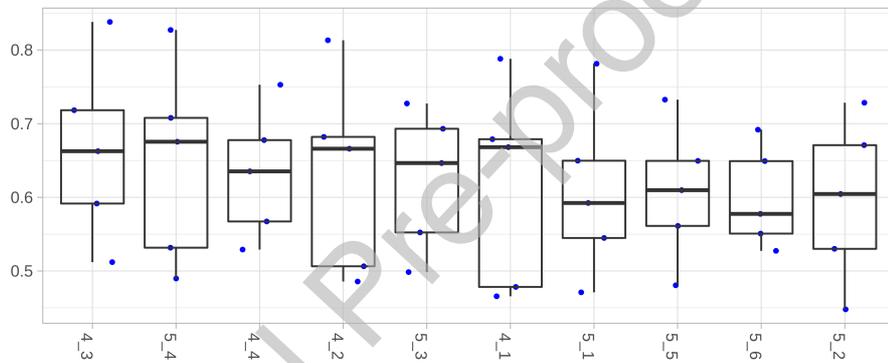


Figure 15: Multi-granularity recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

482 The average AD-Accuracy by model was between 49.1% and 76.8% 16. The
 483 model of NUSControl Lab outperformed the models of the other teams, with a
 484 recognition rate 12 points higher than the second competing team (IMPACT
 485 had a better result than wr0112358 team but was not competing). The team
 486 ranking was not impacted by the ranking method (Figure 17).

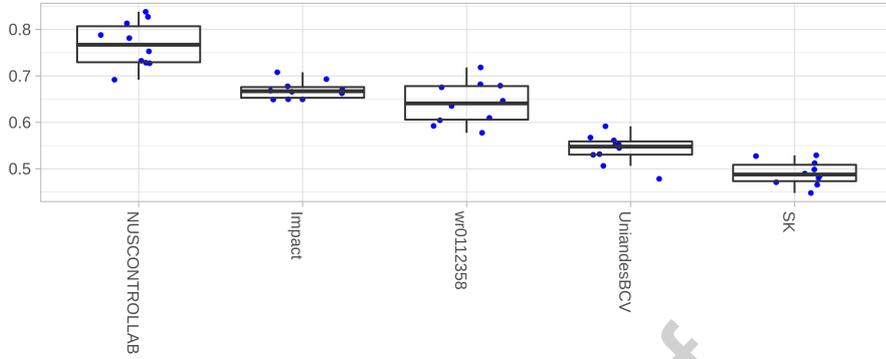


Figure 16: Average multi-granularity recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

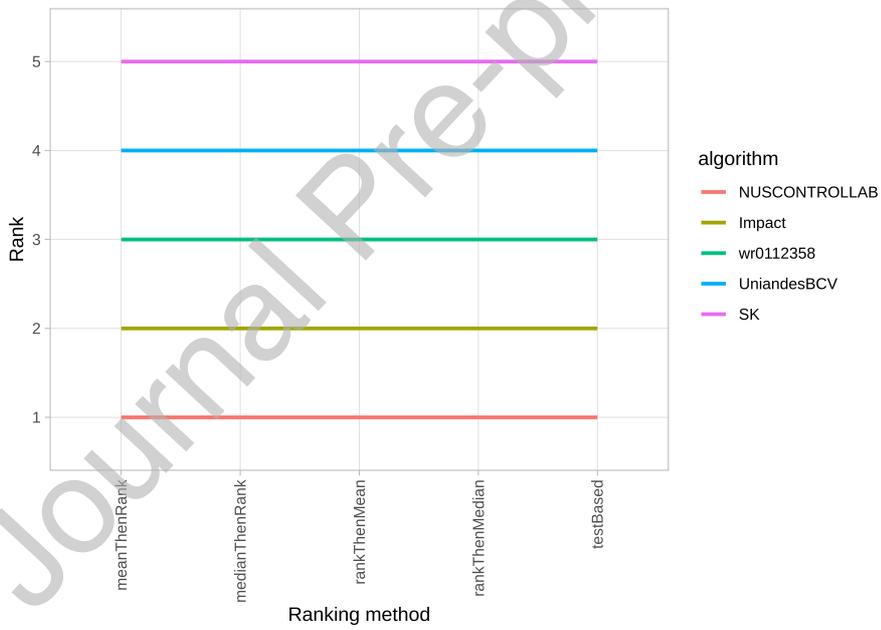


Figure 17: Multi-granularity recognition ranking stability through different ranking methods.

487

Table 4 summarizes the results of each model on all tasks.

Team	Model	Task 1	Task 2	Task 3	Task 4
MedAIR	uni	96.53	84.02		
NUSControl Lab	multi	94.10	74.64	61.69	76.81
SK	multi	58.99	35.85	52.40	49.08
UniandesBCV	uni	89.45	60.21	61.31	
	multi	61.45	39.91	61.08	54.15
Wr0112358	uni	91.60	63.74	58.95	
	multi	84.49	51.41	56.71	64.21
IMPACT	uni	80.66	46.48	58.10	
	multi	82.70	57.08	61.06	66.95

Table 4: Average AD-Accuracy of each model for all tasks. On model, uni is for models dedicated to uni-granularity recognition, multi for models able to recognize all modalities. Best results are highlighted in bold for each task.

488 4. Discussion

489 Surgical workflow recognition is an important challenge in providing automatic
490 context-aware computer-assisted surgical systems. However, as demonstrated by
491 the different models proposed in this challenge, there remains a lot of room for
492 improvement. For a high level of granularity (phases and steps), the best models
493 have a recognition rate that may be sufficient for applications such as prediction
494 of remaining surgical time or resource management. However, for activities, the
495 recognition rates are still insufficient to propose clinical applications.

496 For all tasks, the decrease between the sequence with the best recognition
497 rate and the one with the lowest was linear. The difference between the best
498 and the worst recognition rate was 7 points for phase recognition, 13.2 for step
499 recognition, 8.3 for activity recognition, and 6.8 for multi-granularity recognition.
500 Only sequence 5_6 for phase recognition presented a recognition gap of 3 points
501 with the penultimate sequence (Figure 6). After a review of this sequence,
502 the major difference was a high representation of the "idle" phase (7.13% for
503 sequence 5_6 compared to $2.49 \pm 1.22\%$ for the other test cases) to the detriment

504 of the "suturing" phase (36.12% compared to $45.14 \pm 10.36\%$) for a similar total
505 duration (79s vs. $99s \pm 52s$). However, this cannot be the only reason for this low
506 recognition rate. In the future, it could be interesting to study the explainability
507 of the different networks.

508 For the image modality, all teams proposed a model based on convolutional
509 neural networks (CNNs) such as ResNet, and VGG, two of them also combined
510 a recurrent neural network (RNN) as LSTM. For the kinematic modality, two
511 teams used CNN, one used RNN, and another used a combination of both. The
512 teams wr0112358 and UniandesBCV did not use this modality. According to
513 the results, the use of RNNs seemed to be more relevant than that of CNNs
514 only. However, both teams that used them also performed post-processing to
515 improve the recognition rate, so it was difficult to evaluate the role of the RNNs
516 and post-processing.

517 For the phases and step recognition tasks, the multi-granularity models had
518 globally worse performances than the uni-granularity models, even for the teams
519 who proposed both models on the same architecture. The only exception was
520 for IMPACT, but the results were quite similar between the team's models. For
521 the activity recognition, it was the opposite: 3 of the 4 top models were multi-
522 granularity ones. Two reasons could explain this fact. First, the activity and
523 multi-granularity models had to recognize multiple components at the same time
524 (6 and 8 components, respectively), whereas the other models only recognized 1
525 component. Second, the majority of activities could only appear on a specific
526 phase or step. For example, the activity consisting of inserting a needle on the
527 right artificial vessel with a needle holder (noted <insert, right artificial vessel,
528 needle holder>) could only appear on the phase "suturing" and, specifically, on
529 the step "suture making". So, the multi-granularity models could learn these
530 relations to improve their performances for activity recognition.

531 One of the most surprising results of the challenge was the similar recognition
532 rate between the video-based models and the multi-modality-based models (using
533 both videos and kinematics). Team wr0112358 reported that the kinematic data
534 did not significantly contribute to the performance of their model. This was

535 confirmed by the ranking of this team (top 3 for the phase and step recognition
536 tasks with a dedicated model and top 3 for the multi-granularity task). The
537 UniandesBCV team also used only the video modality, and also had good ranking,
538 especially for activity recognition, with a tie for first place between both models
539 proposed. However, it is impossible to know whether these results come from
540 the models used by the participants or from the lack of information provided by
541 the kinematic data. A more robust and systematic study would clarify this by
542 the understanding of the models and the contribution of each modality.

543 The first main limitation was the unbalanced distribution of cases by expertise
544 level (11 performed by experts, 16 by engineering students) due to the different
545 number of cases by participants (between 3 and 6 cases). We split the data set
546 to have a similar distribution between the training and test data sets to limit
547 the impact of this unbalanced distribution.

548 The second main limitation was the release of the video and kinematic data
549 of the test cases during the challenge. This choice was dictated by the organizers'
550 lack of knowledge of Docker images and the lack of hardware available when the
551 challenge was proposed to EndoVis and MICCAI. So, we wanted to be able to
552 use the results provided by the participants if necessary. Finally, all results were
553 computed on the organizers' hardware via Docker images. With the test cases
554 release, we first asked unnecessary works to teams; the time spent running the
555 results could have been dedicated to the improvement of the methods. Moreover,
556 this early release could have allowed the participants to make their own manual
557 annotations and use them for the training. Even if these annotations were
558 different than those by the organizer, it opened a breach for undetectable fraud.

559 In addition to confirming the superiority of RNNs compared to CNNs with
560 same post-processing method and studying the impact of each modality, future
561 work could explore more complex networks such as hierarchical models. Indeed,
562 the granularity description is hierarchic (a step belongs to a phase; some activities
563 only appear on specific steps), so this type of model could improve the recognition.
564 Enlarging the data set with more sessions, more modalities, and more sources of
565 data (other systems, virtual reality simulators, real surgeries, etc.) is also being

566 considered for a second version of the MISAW challenge.

567 **Acknowledgements**

568 This work was partially by ImPACT Program of Council for Science, Technology
569 and Innovation, Cabinet Office, Government of Japan.

570 Authors thanks the IRT b<>com for the provision of the software “Surgery
571 Workflow Toolbox [annotate]” , used for this work.

572 **Statements of ethical approval**

573 All procedures performed in studies involving human participants were in
574 accordance with the ethical standards of the institutional and/or national research
575 committee and with the 1964 Helsinki declaration and its later amendments or
576 comparable ethical standards. This articles does not contain patient data.

577 **Conflict of interest statement**

578 The authors declare that they have no conflict of interest.

579 **References**

- 580 [1] P. Jannin, M. Raimbault, X. Morandi, B. Gibaud, Modeling surgical pro-
581 cedures for multimodal image-guided neurosurgery, in: Lecture Notes in
582 Computer Science (including subseries Lecture Notes in Artificial Intelli-
583 gence and Lecture Notes in Bioinformatics), Vol. 2208, Springer Verlag,
584 2001, pp. 565–572. doi:10.1007/3-540-45468-3{_}68.
585 URL <http://idm.univ-rennes1.fr/>
- 586 [2] W. S. Sandberg, B. Daily, M. Egan, J. E. Stahl, J. M. Goldman, R. A.
587 Wiklund, D. Rattner, Deliberate Perioperative Systems Design Improves
588 Operating Room Throughput:, *Anesthesiology* 103 (2) (2005) 406–418.
589 doi:10.1097/0000542-200508000-00025.

- 590 URL <http://files/987/Deliberateperioperativesystemsdesignimprovesoperatingroomthroughput>
591 [pdfhttp://files/986/16052124.html](http://files/986/16052124.html)
- 592 [3] B. Bhatia, T. Oates, Y. Xiao, P. Hu, Real-time identification of operating
593 room state from video, in: Proceedings of the National Conference on
594 Artificial Intelligence, Vol. 2, 2007, pp. 1761–1766.
595 URL [http://files/989/Real-timeidentificationofoperatingroomstatefromvideo.](http://files/989/Real-timeidentificationofoperatingroomstatefromvideo.pdf)
596 [pdf](#)
- 597 [4] A. Hualmé, K. Harada, G. Forestier, M. Mitsuishi, P. Jannin, Sequential
598 surgical signatures in micro-suturing task, International Journal of Computer
599 Assisted Radiology and Surgery 13 (9) (2018) 1419–1428. doi:10.1007/
600 s11548-018-1775-x.
- 601 [5] G. Forestier, L. Riffaud, F. Petitjean, P.-L. Henaux, P. Jannin, Surgical skills:
602 Can learning curves be computed from recordings of surgical activities?,
603 International Journal of Computer Assisted Radiology and Surgery 13 (5)
604 (2018) 629–636. doi:10.1007/s11548-018-1713-y.
- 605 [6] S.-Y. Ko, J. Kim, W.-J. Lee, D.-S. Kwon, Surgery task model for intelligent
606 interaction between surgeon and laparoscopic assistant robot, International
607 Journal of Assitive Robotics and Mechatronics 8 (1) (2007) 38–46.
608 URL [http://files/997/2007_SurgeryTaskModelforIntelligentInteractionbetweenSurgeonandLap:](http://files/997/2007_SurgeryTaskModelforIntelligentInteractionbetweenSurgeonandLap.pdf)
609 [pdf](#)
- 610 [7] G. Quellec, M. Lamard, B. Cochener, G. Cazuguel, Real-Time Task Recog-
611 nition in Cataract Surgery Videos Using Adaptive Spatiotemporal Poly-
612 nomials, IEEE Transactions on Medical Imaging 34 (4) (2015) 877–887.
613 doi:10.1109/TMI.2014.2366726.
614 URL <http://ieeexplore.ieee.org/document/6942202/>
- 615 [8] A. Hualmé, P. Jannin, F. Reche, J.-L. Faucheron, A. Moreau-Gaudry,
616 S. Voros, Offline identification of surgical deviations in laparoscopic rec-
617 topexy, Artificial Intelligence in Medicine 104 (2019) 1–26. doi:10.1016/

- 618 j.artmed.2020.101837.
619 URL <http://arxiv.org/abs/1909.10790><http://dx.doi.org/10.1016/>
620 j.artmed.2020.101837
- 621 [9] F. Lalys, P. Jannin, Surgical process modelling: a review, International
622 Journal of Computer Assisted Radiology and Surgery 9 (3) (2013) 495–511.
623 doi:10.1007/s11548-013-0940-5.
624 URL [http://files/940/LalyssetJannin-2013-Surgicalprocessmodellinareview.](http://files/940/LalyssetJannin-2013-Surgicalprocessmodellinareview.pdf)
625 [pdfhttp://files/939/s11548-013-0940-5.html](http://files/939/s11548-013-0940-5.html)
- 626 [10] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet,
627 P. Jannin, Unsupervised trajectory segmentation for surgical gesture recog-
628 nition in robotic training, IEEE Transactions on Biomedical Engineering
629 63 (6) (2015) 1280–1291. doi:10.1109/TBME.2015.2493100. $\{\}$.
630 URL [http://hal-lirmm.ccsd.cnrs.fr/lirmm-01217023/](http://hal-lirmm.ccsd.cnrs.fr/lirmm-01217023/document)
631 [documenthttps://hal-lirmm.ccsd.cnrs.fr/lirmm-01217023](https://hal-lirmm.ccsd.cnrs.fr/lirmm-01217023)
- 632 [11] A. Huauilmé, F. Despinoy, S. A. Heredia Perez, K. Harada, M. Mitsuishi,
633 P. Jannin, Automatic annotation of surgical activities using virtual reality
634 environments, International Journal of Computer Assisted Radiology and
635 Surgery 14 (10) (2019) 1663–1671. doi:10.1007/s11548-019-02008-x.
636 URL [http://www.embase.com/search/results?subaction=](http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L628117494%0A)
637 [viewrecord&from=export&id=L628117494%0Ahttp://dx.doi.](http://dx.doi.org/10.1007/s11548-019-02008-x)
638 [org/10.1007/s11548-019-02008-xhttps://doi.org/10.1007/](https://doi.org/10.1007/s11548-019-02008-x)
639 [s11548-019-02008-x](https://doi.org/10.1007/s11548-019-02008-x)
- 640 [12] O. Zisimopoulos, E. Flouty, M. Stacey, S. Muscroft, P. Giataganas, J. Nehme,
641 A. Chow, D. Stoyanov, Can surgical simulation be used to train detection
642 and classification of neural networks?, in: Healthcare Technology Letters,
643 Vol. 4, 2017, pp. 216–222. doi:10.1049/htl.2017.0064.
- 644 [13] R. DiPietro, G. D. Hager, Automated Surgical Activity Recognition with
645 One Labeled Sequence (10 2019). doi:10.1007/978-3-030-32254-0{_
646 }51.

- 647 URL <https://link.springer.com/chapter/10.1007/>
 648 978-3-030-32254-0_51
- 649 [14] N. Padoy, T. Blum, S.-A. S. A. S.-A. S. A. Ahmadi, H. Feussner, M. O.
 650 M.-O. M. O. M.-O. Berger, N. Navab, Statistical modeling and recognition
 651 of surgical workflow, *Medical Image Analysis* 16 (3) (2010) 632–641.
 652 doi:10.1016/j.media.2010.10.001.
 653 URL [http://files/822/Statisticalmodelingandrecognitionofsurgicalworkflow.](http://files/822/Statisticalmodelingandrecognitionofsurgicalworkflow.pdf)
 654 pdf[http://files/825/Padoyetal.-2012-Statisticalmodelingandrecognitionofsurgicalw.](http://files/825/Padoyetal.-2012-Statisticalmodelingandrecognitionofsurgicalw.pdf)
 655 pdf<http://files/824/S1361841510001131.html><http://dx.doi.org/>
 656 10.1016/j.media.2010.10.
- 657 [15] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin,
 658 N. Padoy, EndoNet: A Deep Architecture for Recognition Tasks on Laparo-
 659 scopic Videos, *IEEE Transactions on Medical Imaging* 36 (1) (2016) 86–97.
 660 doi:10.1109/TMI.2016.2593957.
 661 URL <http://ieeexplore.ieee.org/document/7519080>/[http://arxiv.](http://arxiv.org/abs/1602.03012)
 662 [org/abs/1602.03012](http://arxiv.org/abs/1602.03012)
- 663 [16] L. Bouarfa, P. P. Jonker, J. Dankelman, Discovery of high-level tasks in
 664 the operating room, *Journal of Biomedical Informatics* 44 (3) (2011) 455–462.
 665 URL [http://files/1002/Discoveryofhigh-leveltasksintheoperatingroom.](http://files/1002/Discoveryofhigh-leveltasksintheoperatingroom.pdf)
 666 pdf
- 667 [17] A. James, D. Vieira, B. Lo, A. Darzi, G.-Z. Yang, Eye-Gaze Driven Surgical
 668 Workflow Segmentation, *Medical Image Computing and Computer-Assisted*
 669 *Intervention MICCAI 2007* (2007) 110–117.
 670 URL [http://link.springer.com/content/pdf/10.1007/](http://link.springer.com/content/pdf/10.1007/978-3-540-75759-7_14)
 671 [http://link.springer.com/chapter/](http://link.springer.com/chapter/10.1007/978-3-540-75759-7_14)
 672 [10.1007/978-3-540-75759-7_14](http://link.springer.com/chapter/10.1007/978-3-540-75759-7_14)[http://link.springer.com/](http://link.springer.com/chapter/10.1007/978-3-540-75759-7_14)
 673 [chapter/10.1007/978-3-540-75759-7_14](http://link.springer.com/chapter/10.1007/978-3-540-75759-7_14)<http://files/990/>
 674 Eye-gazedrivensurgicalworkflowsegme
- 675 [18] F. Lalys, D. Bouget, L. Riffaud, P. Jannin, Automatic knowledge-based

- 676 recognition of low-level tasks in ophthalmological procedures, International
677 Journal of Computer Assisted Radiology and Surgery 8 (1) (2012) 39–49.
678 doi:10.1007/s11548-012-0685-6.
679 URL <http://files/1011/IJCARS-FL-2012-fin.pdf>
- 680 [19] L. Zappella, B. Béjar, G. Hager, R. Vidal, Surgical gesture classification from
681 video and kinematic data, Medical Image Analysis 17 (7) (2013) 732–745.
682 doi:10.1016/j.media.2013.04.007.
- 683 [20] D. Sarikaya, P. Jannin, Surgical Gesture Recognition with Optical Flow
684 only, arXiv.
685 URL <http://arxiv.org/abs/1904.01143>
- 686 [21] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, S. Speidel,
687 Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features
688 for Automatic Surgical Gesture Recognition in Video, in: Lecture Notes in
689 Computer Science (including subseries Lecture Notes in Artificial Intelligence
690 and Lecture Notes in Bioinformatics), Vol. 11768 LNCS, 2019, pp. 467–475.
691 doi:10.1007/978-3-030-32254-0{_}52.
692 URL https://gitlab.com/nct_tso_
- 693 [22] J. M. Beer, A. D. Fisk, W. A. Rogers, Toward a Framework for Levels of
694 Robot Autonomy in Human-Robot Interaction, Journal of Human-Robot
695 Interaction 3 (2) (2014) 74. doi:10.5898/jhri.3.2.beer.
- 696 [23] M. Mitsuishi, A. Morita, N. Sugita, S. Sora, R. Mochizuki, K. Tanimoto,
697 Y. M. Baek, H. Takahashi, K. Harada, Master-slave robotic platform and
698 its feasibility study for micro-neurosurgery: Master-slave robotic platform
699 for microneurosurgery, The International Journal of Medical Robotics and
700 Computer Assisted Surgery 9 (2) (2013) 180–189.
- 701 [24] C. Garraud, B. Gibaud, C. Penet, G. Gazuguel, G. Dardenne, P. Jannin, An
702 Ontology-based Software Suite for the Analysis of Surgical Process Model.,
703 in: Proceedings of Surgetica’2014, Chambery, France, 2014, pp. 243–245.

- 704 [25] O. Dergachyova, D. Bouget, A. Huaultmé, X. Morandi, P. Jannin, O. Der-
705 gachyova, D. Bouget, A. Huaultmé, X. Morandi, P. Jannin, X. Morandi
706 CHU Rennes, Automatic data-driven real-time segmentation and recognition
707 of surgical workflow, *International Journal of Computer Assisted Radiology
708 and Surgery* 11 (6) (2016) 1081–1089. doi:10.1007/s11548-016-1371-x.
709 URL <https://hal.archives-ouvertes.fr/hal-01299344>
- 710 [26] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz,
711 T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, C. Feldmann, A. F. Frangi,
712 P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B. A.
713 Landman, K. März, O. Maier, K. Maier-Hein, B. H. Menze, H. Müller, P. F.
714 Neher, W. Niessen, N. Rajpoot, G. C. Sharp, K. Sirinukunwattana, S. Spei-
715 del, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.-W. Wang, M.-
716 A. Weber, G. Zheng, P. Jannin, A. Kopp-Schneider, Why rankings of biomed-
717 ical image analysis competitions should be interpreted with care, *Nature
718 Communications* 9 (1) (2018) 5217. doi:10.1038/s41467-018-07619-7.
719 URL <http://www.nature.com/articles/s41467-018-07619-7>
720 <https://www.nature.com/articles/s41467-018-07619-7.pdf>
- 721 [27] M. Wiesenfarth, A. Reinke, B. A. Landman, M. Eisenmann, L. A. Saiz,
722 M. J. Cardoso, L. Maier-Hein, A. Kopp-Schneider, Methods and open-source
723 toolkit for analyzing and visualizing challenge results, *Scientific Reports*
724 11 (1) (2021) 2369. doi:10.1038/s41598-021-82017-6.
725 URL <https://doi.org/10.1038/s41598-021-82017-6>
- 726 [28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition,
727 in: *Proceedings of the IEEE Computer Society Conference on Computer
728 Vision and Pattern Recognition*, Vol. 2016-Decem, 2016, pp. 770–778. doi:
729 10.1109/CVPR.2016.90.
730 URL <http://image-net.org/challenges/LSVRC/2015/>
- 731 [29] C. Lea, A. Reiter, R. Vidal, G. D. Hager, Segmental spatiotemporal CNNs
732 for fine-grained action segmentation, in: *Lecture Notes in Computer Science*

- 733 (including subseries Lecture Notes in Artificial Intelligence and Lecture
734 Notes in Bioinformatics), Vol. 9907 LNCS, Springer Verlag, 2016, pp. 36–52.
735 doi:10.1007/978-3-319-46487-9{_}3.
736 URL [https://link.springer.com/chapter/10.1007/
737 978-3-319-46487-9_3](https://link.springer.com/chapter/10.1007/978-3-319-46487-9_3)
- 738 [30] R. Dipietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee,
739 M. R. Lee, G. D. Hager, Recognizing surgical activities with recurrent
740 neural networks, in: Lecture Notes in Computer Science (including
741 subseries Lecture Notes in Artificial Intelligence and Lecture Notes in
742 Bioinformatics), Vol. 9900 LNCS, Springer Verlag, 2016, pp. 551–558.
743 doi:10.1007/978-3-319-46720-7{_}64.
744 URL [https://link.springer.com/chapter/10.1007/
745 978-3-319-46720-7_64](https://link.springer.com/chapter/10.1007/978-3-319-46720-7_64)
- 746 [31] Y.-H. Long, J.-Y. Wu, B. Lu, Y.-M. Jin, M. Unberath, Y.-H. Liu, P.-A. Heng,
747 Q. Dou, Relational Graph Learning on Visual and Kinematics Embeddings
748 for Accurate Gesture Recognition in Robotic Surgery, arXiv.
749 URL <http://arxiv.org/abs/2011.01619>
- 750 [32] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C. W. Fu, P. A. Heng, SV-RCNet:
751 Workflow recognition from surgical videos using recurrent convolutional
752 network, IEEE Transactions on Medical Imaging 37 (5) (2018) 1114–1126.
753 doi:10.1109/TMI.2017.2787657.
- 754 [33] M. Tan, Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional
755 Neural Networks, 36th International Conference on Machine Learning, ICML
756 2019 2019-June (2019) 10691–10700.
757 URL <http://arxiv.org/abs/1905.11946>
- 758 [34] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C. W. Fu, P. A. Heng, Multi-task
759 recurrent convolutional network with correlation loss for surgical video
760 analysis, Medical Image Analysis 59 (2020) 101572. doi:10.1016/j.media.
761 2019.101572.

- 762 [35] M. R. Zhang, J. Lucas, G. Hinton, J. Ba, Lookahead Optimizer: k steps
763 forward, 1 step back, Tech. rep. (2019).
764 URL <https://github.com/michaelrzhang/lookahead>.
- 765 [36] C. Feichtenhofer, H. Fan, J. Malik, K. He, SlowFast Networks for Video
766 Recognition, in: Proceedings of the IEEE international conference on com-
767 puter vision, 2019, p. 6202–6211.
768 URL <https://github.com/>
- 769 [37] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan,
770 F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The
771 Kinetics Human Action Video Dataset, arXiv.
772 URL <http://arxiv.org/abs/1705.06950>
- 773 [38] D. Chung, K. Tahboub, E. J. Delp, A Two Stream Siamese Convolutional
774 Neural Network For Person Re-Identification, Tech. rep. (2017).
- 775 [39] Z. Zhang, P. Luo, C. C. Loy, X. Tang, Facial landmark detection by deep
776 multi-task learning, in: Lecture Notes in Computer Science (including
777 subseries Lecture Notes in Artificial Intelligence and Lecture Notes in
778 Bioinformatics), Vol. 8694 LNCS, Springer Verlag, 2014, pp. 94–108.
779 doi:10.1007/978-3-319-10599-4{_}7.
780 URL [https://link.springer.com/chapter/10.1007/
781 978-3-319-10599-4_7](https://link.springer.com/chapter/10.1007/978-3-319-10599-4_7)
- 782 [40] K. Simonyan, A. Zisserman, Two-stream convolutional networks for ac-
783 tion recognition in videos, in: Advances in Neural Information Processing
784 Systems, Vol. 1, 2014, pp. 568–576.
- 785 [41] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P. A. Muller, Evalu-
786 ating Surgical Skills from Kinematic Data Using Convolutional Neural Net-
787 works, in: Lecture Notes in Computer Science (including subseries Lecture
788 Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol.
789 11073 LNCS, 2018, pp. 214–221. doi:10.1007/978-3-030-00937-3{_
790 }25.