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.

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

MIcro-Surgical Anastomose Workflow recognition challenge report

Arnaud Huaulmé[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 Reiter[i], Manoru Mitsuishi[j], Kanako Harada[j], Pierre Jannin[a,*]

[a] *Univ Rennes,INSERM, LTSI - UMR 1099, F35000, Rennes, France*
[b] *Gazi University, Faculty of Engineering; Department of Computer Engineering, Ankara, Turkey*
[c] *Department of Computer Science & Engineering, The Chinese University of Hong Kong*
[d] *T Stone Robotics Institute, The Chinese University of Hong Kong*
[e] *National University of Singapore(NUS), Singapore, Singapore.*
[f] *Southern University of Science and Technology (SUSTech), Shenzhen, China.*
[g] *Konica Minolta, Inc*
[h] *Center for Research and Formation in Artificial Intelligence, Department of Biomedical Engineering, Universidad de los Andes, Bogotá, Colombia*
[i] *Wintegral GmbH*
[j] *Department 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.huaulme@univ-rennes1.fr (Arnaud Huaulmé), 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. Introduction

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

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

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

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

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

The "MIcro-Surgical Anastomose Workflow recognition on training sessions" (MISAW) challenge provided a unique data set for online automatic recognition of multi-granularity surgical workflows using kinematic and stereoscopic video information on a micro-anastomosis training task. The participants were challenged to develop uni-granularity (with phases, steps, or activities) and/ or multi-granularity workflow recognition models.

2. Methods: Reporting of Challenge Design

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

2.1. Challenge organization

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

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

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

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

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

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

2.2. Mission of the challenge

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

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

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

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

2.3. Challenge data set

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

	Training cases				Test cases	
Participant	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

2.3.1. Data acquisition

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

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

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

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.

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

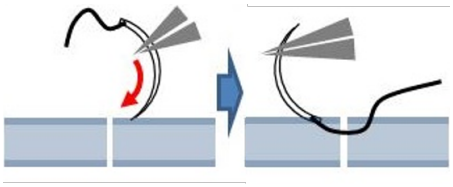


Figure 1: Representation of the beginning (left) and the end (right) of the "suture making" step.

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

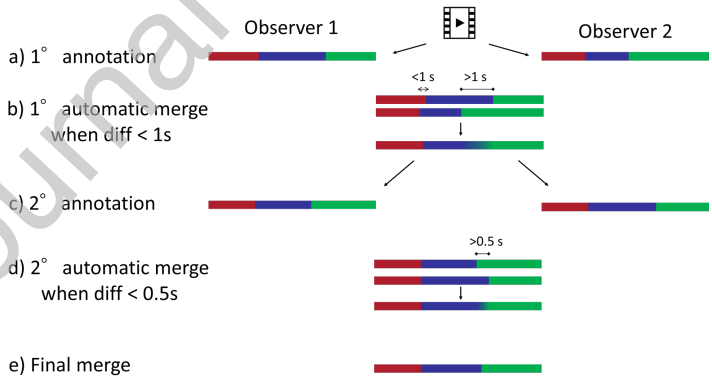
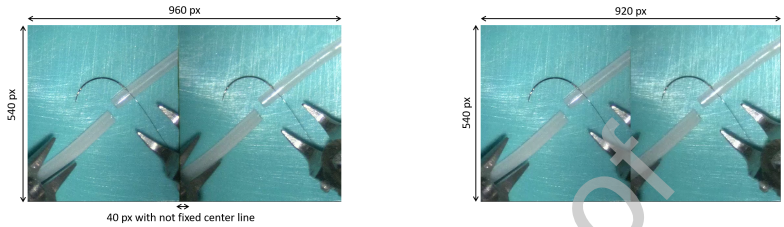


Figure 2: Harmonization protocol used to merge the annotations made by two observers.

2.3.2. Data pre-processing

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

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



(a) (b)

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

.

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

2.3.3. Sources of errors

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

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

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

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

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

2.4. Assessment method

2.4.1. Metrics

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

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

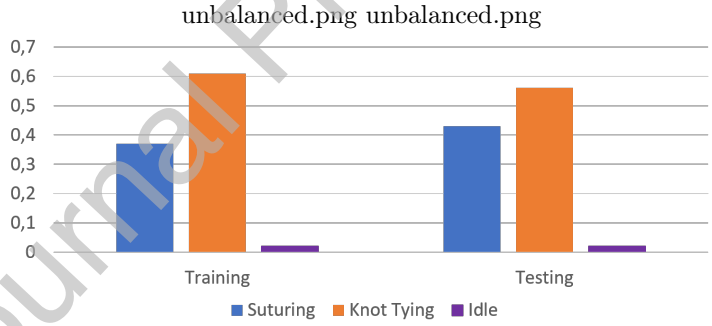


Figure 4: Phase distribution in training and testing data sets

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

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

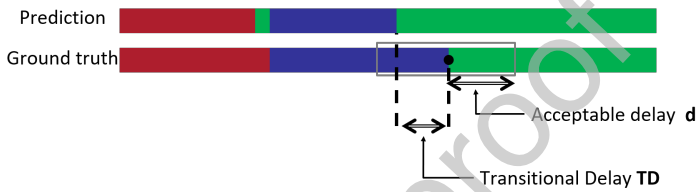


Figure 5: Definition of acceptable and transitional delay used to compute the average-dependent scores

2.4.2. Ranking method

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

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

$$s_{uni}(a_i) = \frac{\sum_{t=0}^{T} balance\_accuracy\_case\_t}{T} \tag{3}$$

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

with each component (i.e., $s_{verb\_Left}(a_i)$) computed with Equation 3:

$$s_{activity}(a_i) = \frac{1}{6} * (s_{verb\_Left}(a_i) + s_{Target\_Left}(a_i) + s_{Instrument\_Left}(a_i) + s_{verb\_Right}(a_i) + s_{Target\_Right}(a_i) + s_{Instrument\_Right}(a_i)) \quad (4)$$

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

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

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

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

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

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

3. Results: Reporting of the Challenge Outcomes

3.1. Challenge submission

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

3.2. Information on participating teams and corresponding methods

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

3.2.1. MedAIR

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

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

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

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

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

3.2.2. NUSControl Lab

The NUSControl Lab team was composed of Chin Boon Chng[1], Wenjun Lin[1,2], Jiaqi Zhang[1], Yaxin Hu[1], Yan Hu[1], Liu Jiang Jimmy[2], and Chee Kong Chui [1]. The participants noted with "1" were from the National University of Singapore (NUS), Singapore, Singapore; the participants noted with "2" were from the Southern University of Science and Technology, Shenzhen, China. This team participated in the multi-granularity task. As described in the subsection "Ranking method" (2.4.2), the model was also ranked in each uni-granularity task.

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

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

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

3.2.3. SK

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

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

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

3.2.4. UniandesBCV

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

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

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

3.2.5. Wr0112358

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

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

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

3.2.6. IMPACT

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

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

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

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

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

3.3. Workflow recognition results

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

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

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

3.3.1. Task 1: Phase recognition

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

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

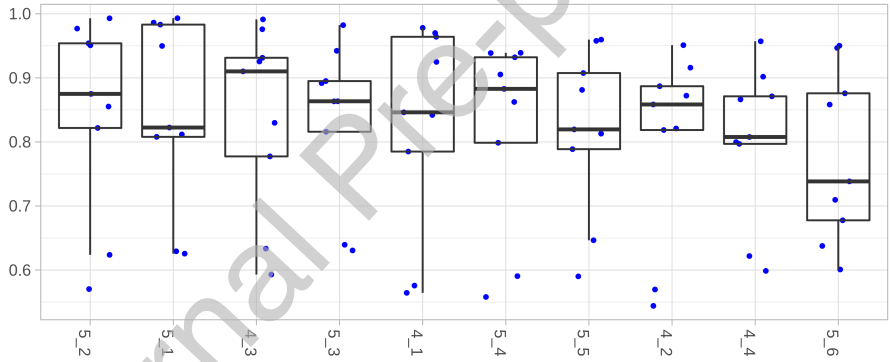


Figure 6: Phase recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

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

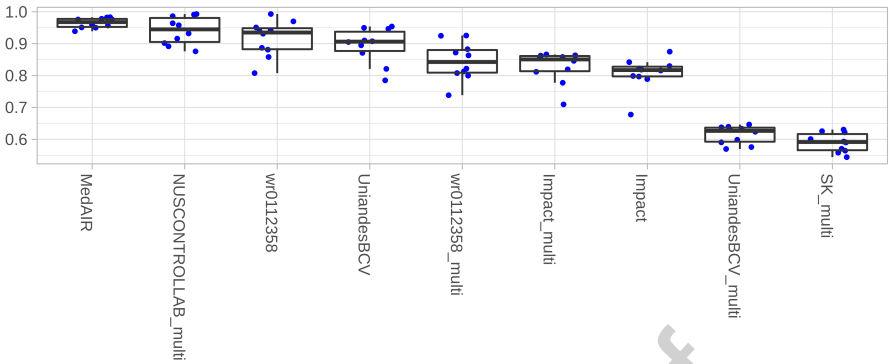


Figure 7: Average phase recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

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

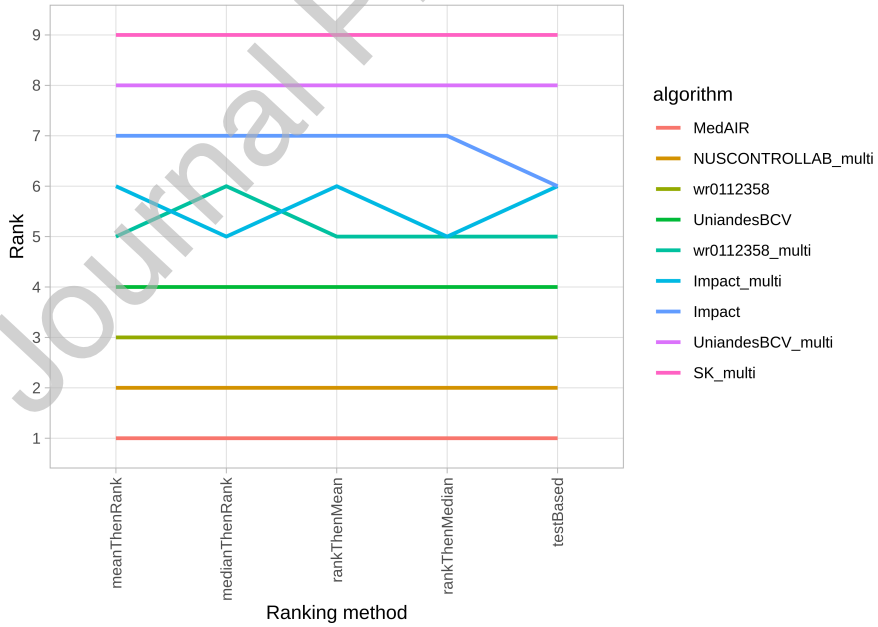


Figure 8: Phase recognition ranking stability through different ranking methods.

3.3.2. Task 2: Step recognition

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

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

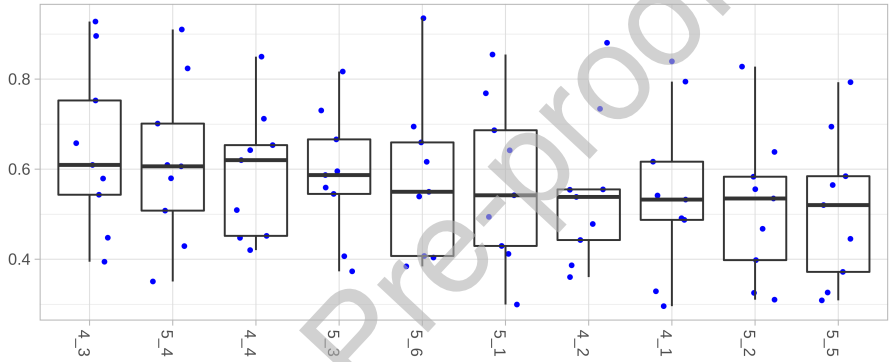


Figure 9: Step recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

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

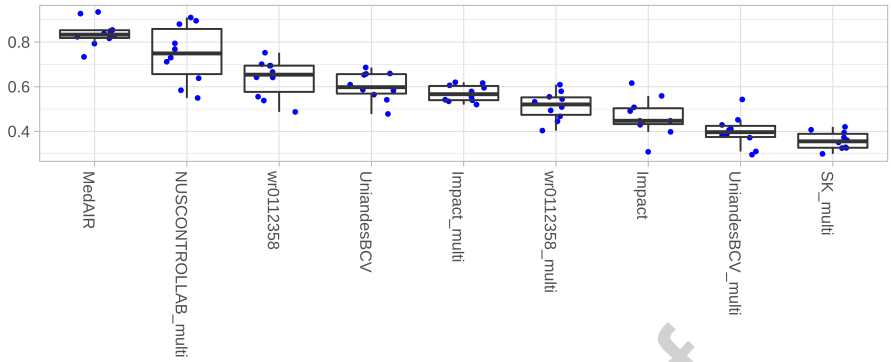


Figure 10: Average step recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

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

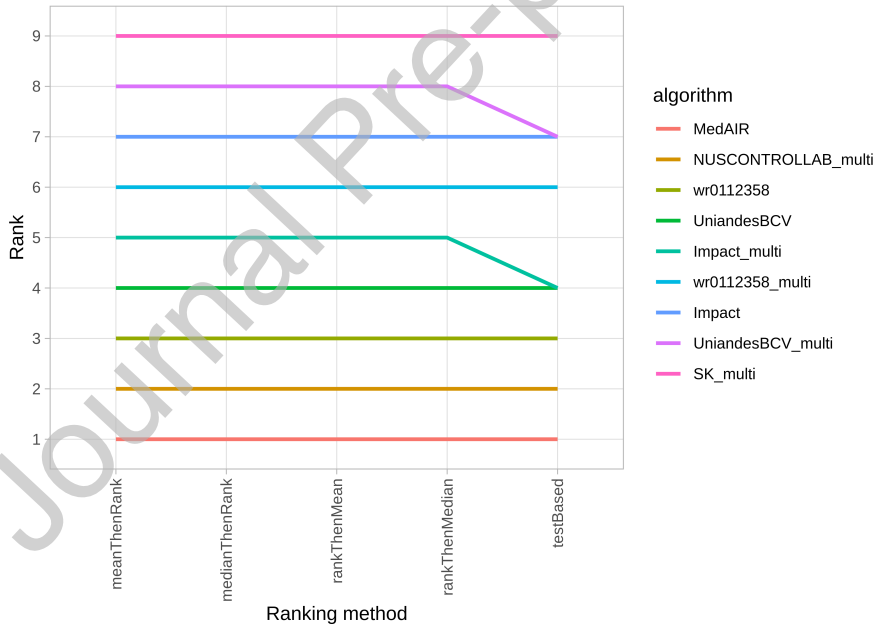


Figure 11: Step recognition ranking stability through different ranking methods.

3.3.3. Task 3: Activity recognition

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

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

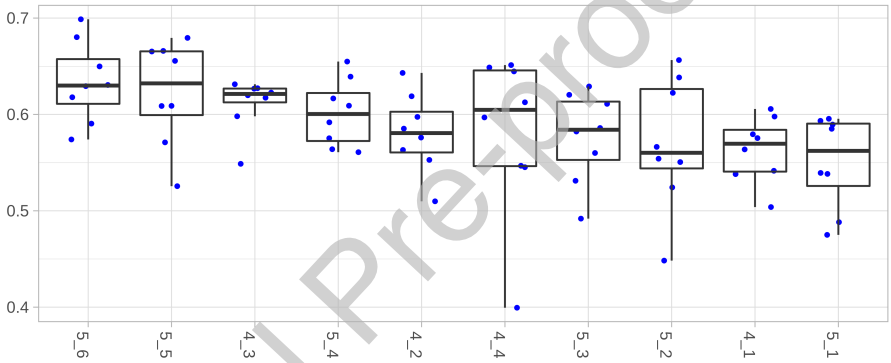


Figure 12: Activity recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

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

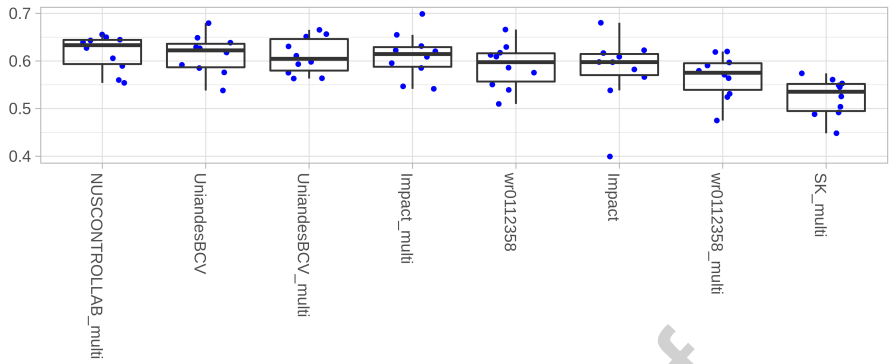


Figure 13: Average activity recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

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

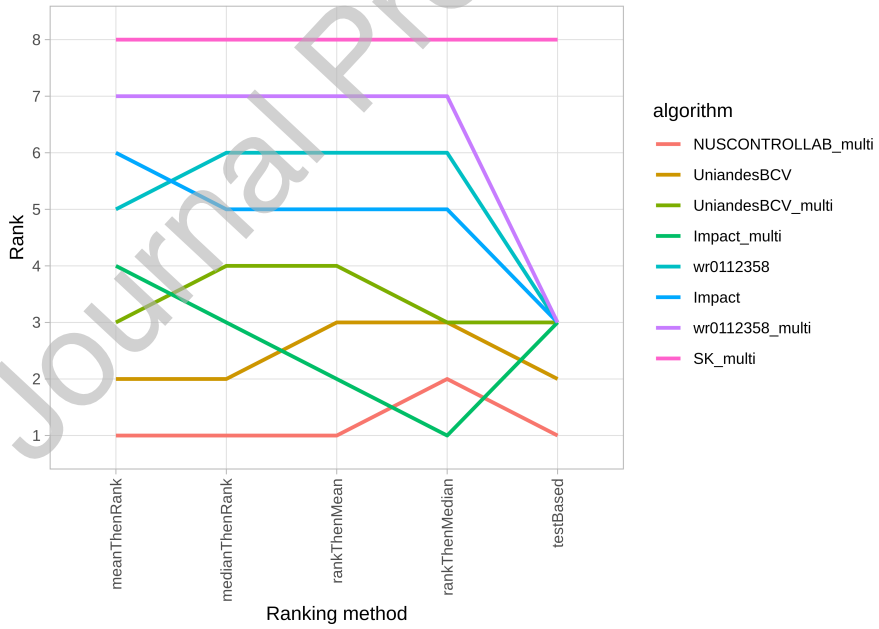


Figure 14: Phase recognition ranking stability through different ranking methods.

3.3.4. Task 4: Multi-granularity recognition

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

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

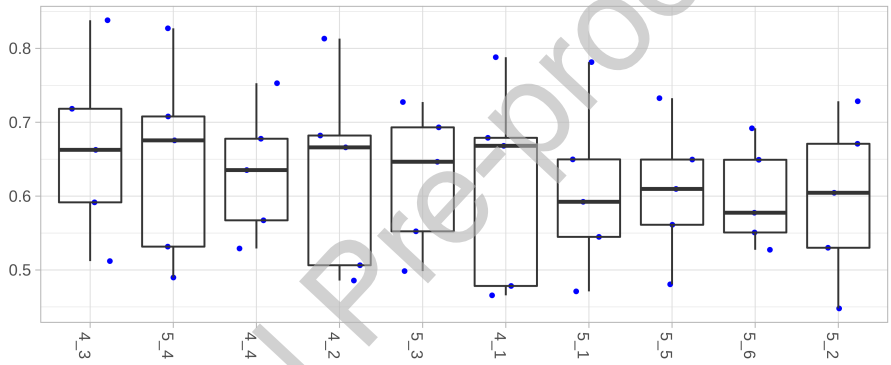


Figure 15: Multi-granularity recognition AD-Accuracy for each sequence. Each dot represents the AD-Accuracy for one model.

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

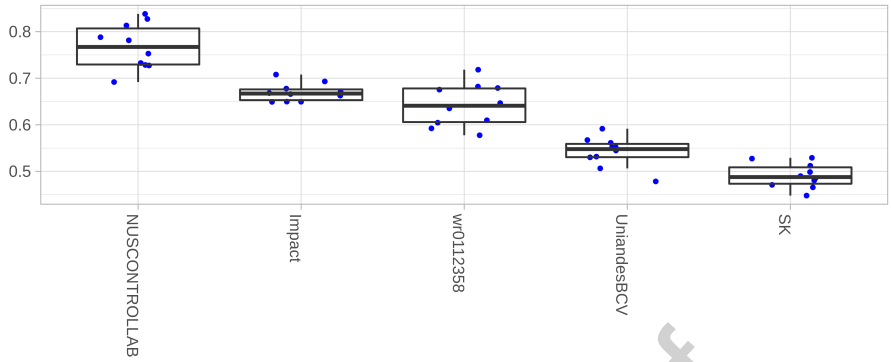


Figure 16: Average multi-granularity recognition AD-Accuracy for each model. Each dot represents the AD-Accuracy for one sequence.

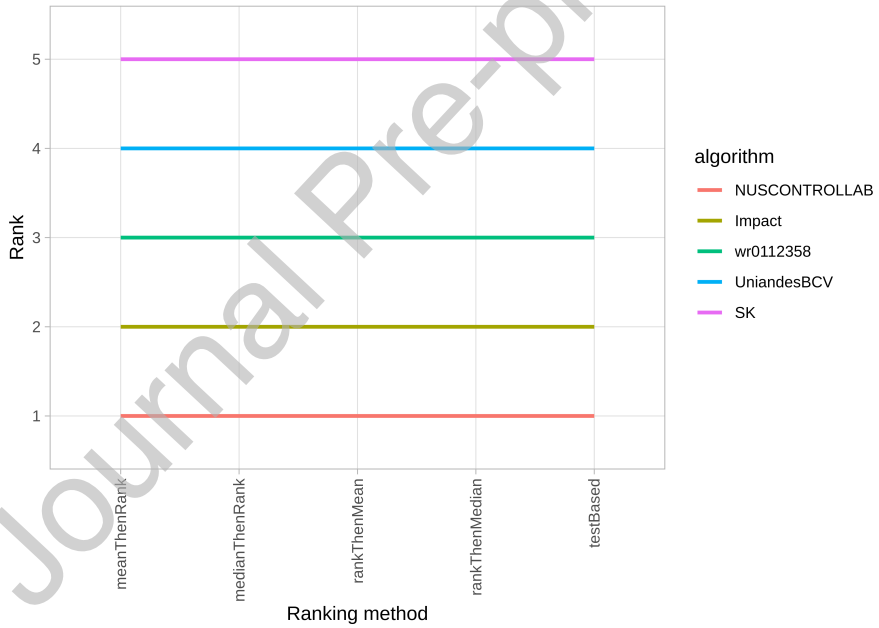


Figure 17: Multi-granularity recognition ranking stability through different ranking methods.

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.

4. Discussion

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

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

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

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

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

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

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

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

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

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

considered for a second version of the MISAW challenge.

Acknowledgements

This work was partially by ImPACT Program of Council for Science, Technology and Innovation, Cabinet Office, Government of Japan.
Authors thanks the IRT b<>com for the provision of the software "Surgery Workflow Toolbox [annotate]" , used for this work.

Statements of ethical approval

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

Conflict of interest statement

The authors declare that they have no conflict of interest.

References

[1] P. Jannin, M. Raimbault, X. Morandi, B. Gibaud, Modeling surgical procedures for multimodal image-guided neurosurgery, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 2208, Springer Verlag, 2001, pp. 565–572. doi:10.1007/3-540-45468-3{_}68.
URL http///idm.univ-rennes1.fr%7C%7C%7C%7C%7C%7C%7C%7C

[2] W. S. Sandberg, B. Daily, M. Egan, J. E. Stahl, J. M. Goldman, R. A. Wiklund, D. Rattner, Deliberate Perioperative Systems Design Improves Operating Room Throughput:, Anesthesiology 103 (2) (2005) 406–418. doi:10.1097/00000542-200508000-00025.

URL http://files/987/Deliberateperioperativesystemsdesignimprovesoperatingroomthroughput
pdfhttp://files/986/16052124.html

[3] B. Bhatia, T. Oates, Y. Xiao, P. Hu, Real-time identification of operating room state from video, in: Proceedings of the National Conference on Artificial Intelligence, Vol. 2, 2007, pp. 1761–1766.
URL http://files/989/Real-timeidentificationofoperatingroomstatefromvideo.pdf

[4] A. Huaulmé, K. Harada, G. Forestier, M. Mitsuishi, P. Jannin, Sequential surgical signatures in micro-suturing task, International Journal of Computer Assisted Radiology and Surgery 13 (9) (2018) 1419–1428. doi:10.1007/s11548-018-1775-x.

[5] G. Forestier, L. Riffaud, F. Petitjean, P. L. Henaux, P. Jannin, Surgical skills: Can learning curves be computed from recordings of surgical activities?, International Journal of Computer Assisted Radiology and Surgery 13 (5) (2018) 629–636. doi:10.1007/s11548-018-1713-y.

[6] S.-Y. Ko, J. Kim, W.-J. Lee, D.-S. Kwon, Surgery task model for intelligent interaction between surgeon and laparoscopic assistant robot, International Journal of Assitive Robotics and Mechatronics 8 (1) (2007) 38–46.
URL http://files/997/2007_SurgeryTaskModelforIntelligentInteractionbetweenSurgeonandLapa
pdf

[7] G. Quellec, M. Lamard, B. Cochener, G. Cazuguel, Real-Time Task Recognition in Cataract Surgery Videos Using Adaptive Spatiotemporal Polynomials, IEEE Transactions on Medical Imaging 34 (4) (2015) 877–887. doi:10.1109/TMI.2014.2366726.
URL http://ieeexplore.ieee.org/document/6942202/

[8] A. Huaulmé, P. Jannin, F. Reche, J.-L. Faucheron, A. Moreau-Gaudry, S. Voros, Offline identification of surgical deviations in laparoscopic rectopexy, Artificial Intelligence in Medicine 104 (2019) 1–26. doi:10.1016/

j.artmed.2020.101837.
URL http://arxiv.org/abs/1909.10790http://dx.doi.org/10.1016/j.artmed.2020.101837

[9] F. Lalys, P. Jannin, Surgical process modelling: a review, International Journal of Computer Assisted Radiology and Surgery 9 (3) (2013) 495–511. doi:10.1007/s11548-013-0940-5.
URL http://files/940/LalysetJannin-2013-Surgicalprocessmodellingareview.pdfhttp://files/939/s11548-013-0940-5.html

[10] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, P. Jannin, Unsupervised trajectory segmentation for surgical gesture recognition in robotic training, IEEE Transactions on Biomedical Engineering 63 (6) (2015) 1280–1291. doi:10.1109/TBME.2015.2493100{\"{i}}.
URL http://hal-lirmm.ccsd.cnrs.fr/lirmm-01217023/documenthttps://hal-lirmm.ccsd.cnrs.fr/lirmm-01217023

[11] A. Huaulmé, F. Despinoy, S. A. Heredia Perez, K. Harada, M. Mitsuishi, P. Jannin, Automatic annotation of surgical activities using virtual reality environments, International Journal of Computer Assisted Radiology and Surgery 14 (10) (2019) 1663–1671. doi:10.1007/s11548-019-02008-x.
URL http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L628117494%0Ahttp://dx.doi.org/10.1007/s11548-019-02008-xhttps://doi.org/10.1007/s11548-019-02008-x

[12] O. Zisimopoulos, E. Flouty, M. Stacey, S. Muscroft, P. Giataganas, J. Nehme, A. Chow, D. Stoyanov, Can surgical simulation be used to train detection and classification of neural networks?, in: Healthcare Technology Letters, Vol. 4, 2017, pp. 216–222. doi:10.1049/htl.2017.0064.

[13] R. DiPietro, G. D. Hager, Automated Surgical Activity Recognition with One Labeled Sequence (10 2019). doi:10.1007/978-3-030-32254-0{_}51.

URL https://link.springer.com/chapter/10.1007/978-3-030-32254-0_51

[14] N. Padoy, T. Blum, S.-A. S. A. S.-A. S. A. Ahmadi, H. Feussner, M. O. M.-O. M. O. M.-O. Berger, N. Navab, Statistical modeling and recognition of surgical workflow, Medical Image Analysis 16 (3) (2010) 632–641. doi:10.1016/j.media.2010.10.001.
URL http://files/822/Statisticalmodelingandrecognitionofsurgicalworkflow.pdfhttp://files/825/Padoyetal.-2012-Statisticalmodelingandrecognitionofsurgicalw.pdfhttp://files/824/S1361841510001131.htmlhttp://dx.doi.org/10.1016/j.media.2010.10.

[15] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, N. Padoy, EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos, IEEE Transactions on Medical Imaging 36 (1) (2016) 86–97. doi:10.1109/TMI.2016.2593957.
URL http://ieeexplore.ieee.org/document/7519080/http://arxiv.org/abs/1602.03012

[16] L. Bouarfa, P. P. Jonker, J. Dankelman, Discovery of high-level tasks in the operating room, Journal of Biomedical Informatics 44 (3) (2011) 455–462.
URL http://files/1002/Discoveryofhigh-leveltasksintheoperatingroom.pdf

[17] A. James, D. Vieira, B. Lo, A. Darzi, G.-Z. Yang, Eye-Gaze Driven Surgical Workflow Segmentation, Medical Image Computing and Computer-Assisted Intervention MICCAI 2007 (2007) 110–117.
URL http://link.springer.com/content/pdf/10.1007%2F978-3-540-75759-7_14.pdfhttp://link.springer.com/chapter/10.1007/978-3-540-75759-7_14http://link.springer.com/chapter/10.1007/978-3-540-75759-7%5C_14http://files/990/Eye-gazedrivensurgicalworkflowsegme

[18] F. Lalys, D. Bouget, L. Riffaud, P. Jannin, Automatic knowledge-based

recognition of low-level tasks in ophthalmological procedures, International Journal of Computer Assisted Radiology and Surgery 8 (1) (2012) 39–49. doi:10.1007/s11548-012-0685-6.
URL http://files/1011/IJCARS-FL-2012-fin.pdf

[19] L. Zappella, B. Béjar, G. Hager, R. Vidal, Surgical gesture classification from video and kinematic data, Medical Image Analysis 17 (7) (2013) 732–745. doi:10.1016/j.media.2013.04.007.

[20] D. Sarikaya, P. Jannin, Surgical Gesture Recognition with Optical Flow only, arXiv.
URL http://arxiv.org/abs/1904.01143

[21] I. Funke, S. Bodenstedt, F. Oehme, F. von Bechtolsheim, J. Weitz, S. Speidel, Using 3D Convolutional Neural Networks to Learn Spatiotemporal Features for Automatic Surgical Gesture Recognition in Video, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11768 LNCS, 2019, pp. 467–475. doi:10.1007/978-3-030-32254-0{_}52.
URL https://gitlab.com/nct_tso_

[22] J. M. Beer, A. D. Fisk, W. A. Rogers, Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction, Journal of Human-Robot Interaction 3 (2) (2014) 74. doi:10.5898/jhri.3.2.beer.

[23] M. Mitsuishi, A. Morita, N. Sugita, S. Sora, R. Mochizuki, K. Tanimoto, Y. M. Baek, H. Takahashi, K. Harada, Master-slave robotic platform and its feasibility study for micro-neurosurgery: Master-slave robotic platform for microneurosurgery, The International Journal of Medical Robotics and Computer Assisted Surgery 9 (2) (2013) 180–189.

[24] C. Garraud, B. Gibaud, C. Penet, G. Gazuguel, G. Dardenne, P. Jannin, An Ontology-based Software Suite for the Analysis of Surgical Process Model., in: Proceedings of Surgetica'2014, Chambery, France, 2014, pp. 243–245.

[25] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, P. Jannin, O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, P. Jannin, X. Morandi CHU Rennes, Automatic data-driven real-time segmentation and recognition of surgical workflow, International Journal of Computer Assisted Radiology and Surgery 11 (6) (2016) 1081–1089. doi:10.1007/s11548-016-1371-x.
URL https://hal.archives-ouvertes.fr/hal-01299344

[26] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A. P. Bradley, A. Carass, C. Feldmann, A. F. Frangi, P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B. A. Landman, K. März, O. Maier, K. Maier-Hein, B. H. Menze, H. Müller, P. F. Neher, W. Niessen, N. Rajpoot, G. C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, A. Kopp-Schneider, Why rankings of biomedical image analysis competitions should be interpreted with care, Nature Communications 9 (1) (2018) 5217. doi:10.1038/s41467-018-07619-7.
URL http://www.nature.com/articles/s41467-018-07619-7https://www.nature.com/articles/s41467-018-07619-7.pdf

[27] M. Wiesenfarth, A. Reinke, B. A. Landman, M. Eisenmann, L. A. Saiz, M. J. Cardoso, L. Maier-Hein, A. Kopp-Schneider, Methods and open-source toolkit for analyzing and visualizing challenge results, Scientific Reports 11 (1) (2021) 2369. doi:10.1038/s41598-021-82017-6.
URL https://doi.org/10.1038/s41598-021-82017-6

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2016-Decem, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
URL http://image-net.org/challenges/LSVRC/2015/

[29] C. Lea, A. Reiter, R. Vidal, G. D. Hager, Segmental spatiotemporal CNNs for fine-grained action segmentation, in: Lecture Notes in Computer Science

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 9907 LNCS, Springer Verlag, 2016, pp. 36–52. doi:10.1007/978-3-319-46487-9{_}3.
URL https://link.springer.com/chapter/10.1007/978-3-319-46487-9_3

[30] R. Dipietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, G. D. Hager, Recognizing surgical activities with recurrent neural networks, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 9900 LNCS, Springer Verlag, 2016, pp. 551–558. doi:10.1007/978-3-319-46720-7{_}64.
URL https://link.springer.com/chapter/10.1007/978-3-319-46720-7_64

[31] Y.-H. Long, J.-Y. Wu, B. Lu, Y.-M. Jin, M. Unberath, Y.-H. Liu, P.-A. Heng, Q. Dou, Relational Graph Learning on Visual and Kinematics Embeddings for Accurate Gesture Recognition in Robotic Surgery, arXiv.
URL http://arxiv.org/abs/2011.01619

[32] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C. W. Fu, P. A. Heng, SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network, IEEE Transactions on Medical Imaging 37 (5) (2018) 1114–1126. doi:10.1109/TMI.2017.2787657.

[33] M. Tan, Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 36th International Conference on Machine Learning, ICML 2019 2019-June (2019) 10691–10700.
URL http://arxiv.org/abs/1905.11946

[34] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C. W. Fu, P. A. Heng, Multi-task recurrent convolutional network with correlation loss for surgical video analysis, Medical Image Analysis 59 (2020) 101572. doi:10.1016/j.media.2019.101572.

[35] M. R. Zhang, J. Lucas, G. Hinton, J. Ba, Lookahead Optimizer: k steps forward, 1 step back, Tech. rep. (2019).
URL https://github.com/michaelrzhang/lookahead.

[36] C. Feichtenhofer, H. Fan, J. Malik, K. He, SlowFast Networks for Video Recognition, in: Proceedings of the IEEE international conference on computer vision, 2019, p. 6202–6211.
URL https://github.com/

[37] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The Kinetics Human Action Video Dataset, arXiv.
URL http://arxiv.org/abs/1705.06950

[38] D. Chung, K. Tahboub, E. J. Delp, A Two Stream Siamese Convolutional Neural Network For Person Re-Identification, Tech. rep. (2017).

[39] Z. Zhang, P. Luo, C. C. Loy, X. Tang, Facial landmark detection by deep multi-task learning, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 8694 LNCS, Springer Verlag, 2014, pp. 94–108. doi:10.1007/978-3-319-10599-4{_}7.
URL https://link.springer.com/chapter/10.1007/978-3-319-10599-4_7

[40] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Advances in Neural Information Processing Systems, Vol. 1, 2014, pp. 568–576.

[41] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P. A. Muller, Evaluating Surgical Skills from Kinematic Data Using Convolutional Neural Networks, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11073 LNCS, 2018, pp. 214–221. doi:10.1007/978-3-030-00937-3{_}25.