Noname manuscript No.
(will be inserted by the editor)

"Deep-Onto" network for surgical workflow and context recognition

Hirenkumar Nakawala · Roberto Bianchi · Laura Erica Pescatori · Ottavio De Cobelli · Giancarlo Ferrigno · Elena De Momi

Received: date / Accepted: date

Abstract
Purpose Surgical workflow recognition and context-aware systems could allow better decision making and surgical planning by providing the focused information, which may eventually enhance surgical outcomes. While current developments in computer-assisted surgical systems are mostly focused on recognizing surgical phases, they lack recognition of surgical workflow sequence and other contextual element, e.g. "Instruments". Our study proposes a hybrid approach i.e. using deep learning and knowledge representation, to facilitate recognition of the surgical workflow.
Methods We implemented "Deep-Onto" network, which is an ensemble of deep learning models and knowledge management tools, ontology and production rules. As a prototypical scenario, we chose Robot-Assisted Partial Nephrectomy (RAPN). We annotated RAPN videos with surgical entities, e.g. "Step" and so forth. We performed different experiments, including the inter-subject variability, to recognize surgical steps. The corresponding subsequent steps along with other surgical contexts, i.e. "Actions", "Phase" and "Instruments" were also recognized.
Results The system was able to recognize 10 RAPN steps with the prevalence-weighted macro-average (PWMA) recall of 0.83, PWMA precision of 0.74, PWMA F1 score of 0.76, and the accuracy of 74.29% on 9 videos of RAPN.
Conclusion We found that the combined use of deep learning and knowledge representation techniques is a promising approach for the multi-level recognition of RAPN surgical workflow.

H. Nakawala, L. E. Pescatori, G. Ferrigno and E. De Momi
Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133, Milan, Italy.
E-mail: hirenkumar.nakawala@polimi.it

R. Bianchi and O. De Cobelli
Department of Urology, European Institute of Oncology (IEO), Via Giuseppe Ripamonti, 435, Milan, 20141, Italy.

Keywords Deep learning · knowledge representation · Robot-Assisted Partial Nephrectomy · surgical workflow

1 Introduction

In the last decade, a lot of emphasis has been given on the development of surgical assistance systems, by enhancing functionalities of the current regime in Robot-Assisted Surgery (RAS), which could help performing monotonous and simple tasks robustly [1]. Novice surgeons are especially vulnerable to detrimental effects of cognitive overload, due to information overload, causing the preventable adverse events [2]. Besides, because of experience and saved mental models of surgeries, expert surgeons may have gradually developed strategies to cope with the information overload by focusing on the information they need [3]. To automatically recognize a surgical task in progress, i.e. operational steps and sequences, video data processing is an essential step towards context-awareness but a very challenging problem.

As a prototypical scenario, we chose Robotic-Assisted Partial Nephrectomy (RAPN). RAPN regards the removal of a renal tumor. In 2012, estimated prevalence of renal cancer was around 338,000 cases (2% of total cancer cases) in Europe [4]. Surgery is considered as a de-facto treatment for kidney tumor, with 5-year cancer free rates around 95% in large-scale cohorts [5]. However, RAPN has been reported with overall complications ranging from 12.3% to 33% with different surgical approaches, as demonstrated by [6]. RAPN-related adverse events have also been reported [7] such as liver injury, spleen injury, bowel injury, bleeding after vascular clamp removal, renal artery injury, epigastric artery, and renal vein injury, where automatic recognition of workflow would be helpful.

In RAPN, tool-based recognition of surgical workflow, which was widely used e.g. [8] for other surgeries, may not be a practical solution as only three robotic tools are used, i.e. "monopolar curved scissors", "fenestrated bipolar" and "robotic large needle driver", to perform surgical manoeuvres, where additional semantic information, e.g. between tool and actions might be helpful. An ontology provides an explicit specification of concepts within a domain of interest, which could be used to represent "Surgical Process Model" (SPM). In the past, ontologies were used for "phase" recognition in laparoscopic surgeries [9] and context-aware training in percutaneous surgeries [10]. However, perceptual object, e.g. surgical tools in videos, is difficult to be recognized with knowledge-based techniques. Recently, a Convolutional Neural Network (CNN), consisting of 9 layers, was used to extract discriminative feature from the images representing cholecystectomy phases [11]. However, the authors were only focused on recognizing surgical phases without considering the surgical sequence and other semantic information. Current researches [12-13] in the surgical workflow analysis are moving towards recognition of sequences of surgical phases using deep learning. For the workflow recognition, a Convolutional Recurrent Neural Network ("CRNN") and CNN with Hidden Markov

Models (CNN-HMM) were employed on the annotated video data of laparoscopic cholecystectomy. However, due to computational and data limitations, previous researches lack recognition of surgical workflow at multiple levels, i.e. steps, anatomy, instruments and so on.

In this manuscript, we present a pipeline, "Deep-Onto" network, which recognized surgical workflow entities at different granularities, by combining a bottom-up approach, i.e. deep learning, with a top-down approach, i.e. ontologies, thanks to higher expressiveness and semantic relations between surgical entities. The "Deep-Onto" network is an ensemble of two components: 1) a "CRNN" and a "Sequence" model to recognize the surgical step and a subsequent step from RAPN videos; and 2) a "Knowledge" model, which contains an ontology-based SPM on RAPN and logical rules to recognize other surgical context, e.g. instruments. The aim is to automatically understand RAPN workflow, which could be used in a context-aware system framework [10] and eventually assist novice surgeons during surgical training by presenting the contextual information, e.g. the next step during the intervention. As far as our knowledge allows, this is a first implementation of a combined use of deep learning, and knowledge representation and reasoning techniques for the automatic surgical workflow analysis on robot-assisted urological surgery.

2 Methods

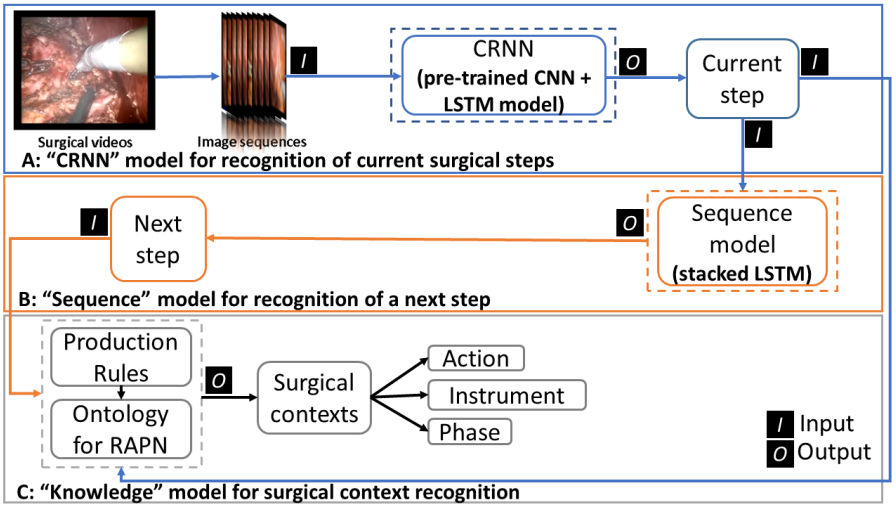


Fig. 1: Proposed "Deep-Onto" Network: schematic diagram, consisting of "CRNN", "Sequence" and "Knowledge" models

The "Deep-Onto" network is shown in Fig. 1. The "CRNN" (subsection 2.1) is used to recognize ongoing surgical step. Thereafter, "Sequence" model

(subsection 2.2) is used to predict the next step based on the current step recognized by "CRNN". The next step predicted by the "Sequence" model is also used as a binary predicate (along with current step) representing step sequence inside the "Knowledge" model. The "Knowledge" model (subsection 2.3) takes input of the predictions of the "CRNN" model i.e. current step, the "Sequence" model i.e. subsequent step, and anatomical information, which is explicitly grounded for each step, to derive other contextual information of the current step i.e. "Phase", "Instrument", and "Actions". The represented pipeline is yet not trained end-to-end.

2.1 "CRNN" model

A combination of CNN and Long Short Term Memory (LSTM) units, a "LRCN" or "CRNN" model [14], has provided excellent results on the video classification tasks e.g. actions. In our modified "CRNN" model, we used Inception V3 [15] as a CNN model, pre-trained on the ImageNet dataset [16]. As shown in Fig. 2, first, Inception V3 is fine-tuned on 10 classes of RAPN steps, i.e. "mobilization", and so on. The final classification layer was removed and 2048-dimensional feature vectors were extracted from Inception V3's global average pooling layer (GAP).

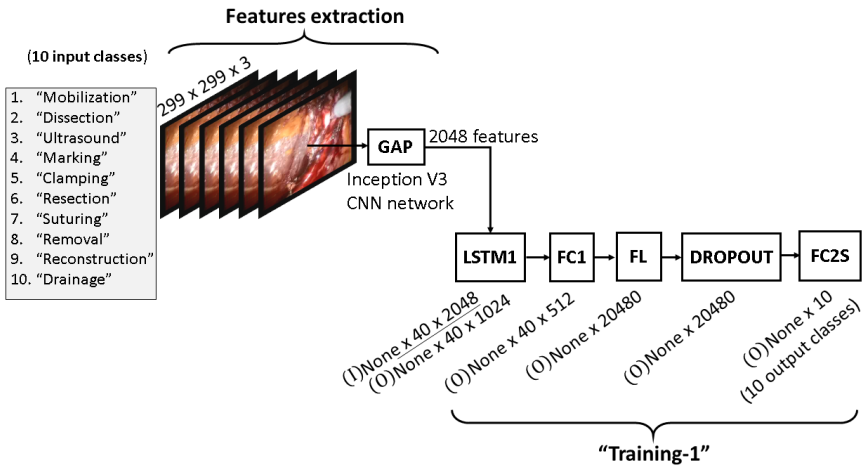


Fig. 2: "CRNN" model configuration during the network training("I" represents input tensor and "O" represents output tensor. "None" represents the batch size, which was 32)

For "CRNN" training, "Training-1", we used every 40th frame in the sequence of frames, i.e. videos for each surgical step, which is chosen empirically based on the best accuracy on the validation set. Then, the extracted

feature vectors, representing videos of 10 steps, is passed to a separate one-layer LSTM network (LSTM1) with 1024 hidden nodes, followed by "Fully-connected" layer (FC1) with 512 hidden nodes. We then used a "Flatten" layer (FL) to map input shape to the 1-dimensional tensor, followed by a "dropout" layer (DROPOUT) of the same output shape and a "Fully-connected" layer with softmax activation function for classification of 10 steps (FC2S).

We implemented the adaptive learning rate, which was reduced to 50% if the accuracy of validation set stopped improving at every three epochs. We used a low initial learning rate, i.e. $1 \times e^{-4}$, to update the network parameters. As regularization methods, i.e. to minimize over-fitting, we used dropout units with the value of 0.5, i.e. DROPOUT in Fig. 2, and early stopping of the training, if the loss function was not improved for 20 epochs.

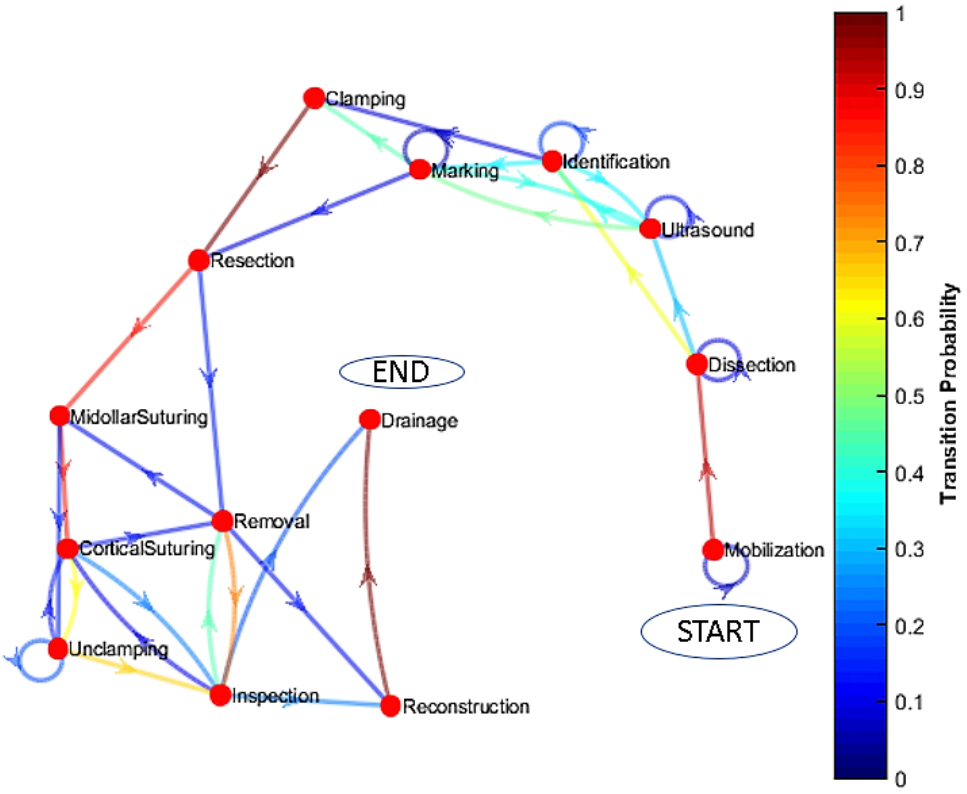
The network was trained using Adam [17] optimizer with categorical cross-entropy (log loss), H, as a loss function as shown in Eq. 1, where p is a set of true labels, representing surgical steps, and q is a set of predicted labels, which contains probabilities obtained from the softmax classification layer. i represents the class index.

$$H(p, q) = -\sum_{i} p_i \log q_i \tag{1}$$

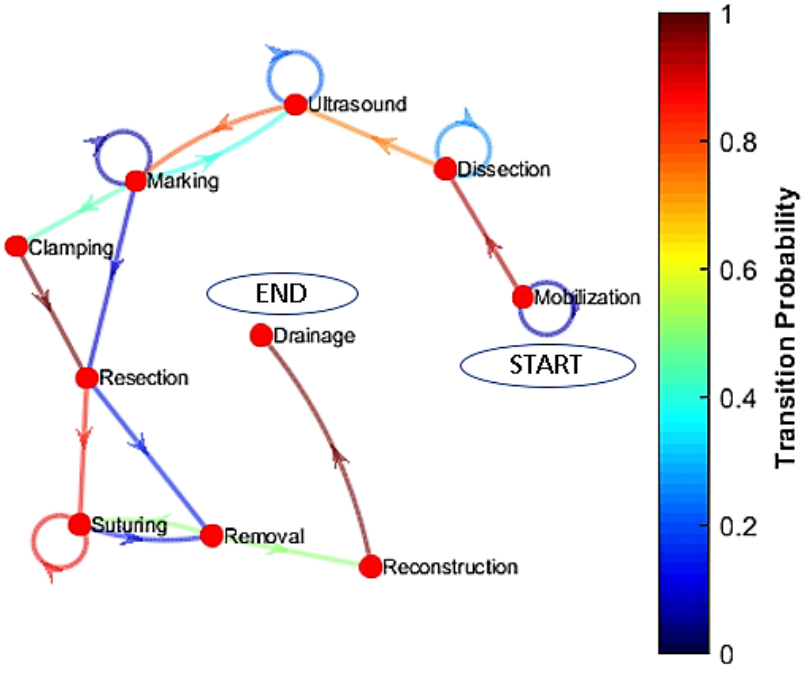
2.2 "Sequence" model

The "Sequence" model was used to predict surgical step sequences. First, as a training set, we generated 1000 random input-output pairs of words representing step sequences. As shown in Fig. 4, these pairs were fed into stacked LSTM, consisting in layers with 32 (LSTM2) and 16 hidden nodes (LSTM3), followed by a "Fully-connected" layer (FC3S) with the softmax activation function for the classification of 10 step sequences. To constraint model's predictive capability to only one consecutive step, size of the sequence length was kept 1 in the training set. The "Sequence" model was trained ("Training-2") using the same methodology, i.e. using Adam optimizer with categorical cross-entropy as a loss function, as explained in subsection 2.1. A predicted step sequence is also used as instances of an ontological relation "hasNextStep", which specifies a step in the progress and a consecutive step (see subsection 2.3).

As shown in Fig. 3, we built a Discrete-Time Markov Chain (DTMC), as a model to obtain the most probable RAPN step sequences, from a step transition matrix obtained by analyzing transitions between and within the steps in 9 video annotations. However, in this manuscript, we only considered the step sequences with highest transition probabilities to form a hierarchical RAPN workflow. Because of the hierarchical step sequences, the LSTM units in "Deep-Onto" network could be replaced by a simple look-up table (see subsections 3.2 and 4.2). However, we considered LSTM units into the pipeline because it does not deteriorate the results and helps in order to prepare for future experiments with more complex transition data.



(a) Model obtained with video annotations



(b) Presented model in our work

Fig. 3: Figure (a) shows the state transition diagram, where each state represents steps of RAPN as in video annotations, starting from "Mobilization". Figure (b) shows the considered step transition model in our work. The color of edges represents transition probabilities between one step to one or more steps.

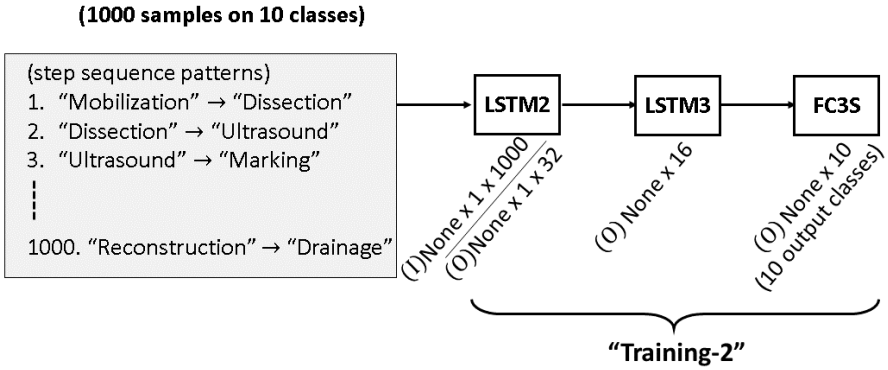


Fig. 4: "Sequence" model configuration during the network training ("I" represents input tensor and "O" represents output tensor. "None" represents the batch size, which was 1)

2.3 "Knowledge" model

A "RAPN ontology" was built using a top-down approach, where the most general concepts of the domain, such as phases (e.g. "hilumDissection") were first analyzed and then specialized concepts, such as actions (e.g. "cut"), were implemented. The needed information about RAPN was obtained from a journal article [18], video annotations and in close collaboration with a urologist ("RB").

Logical sentences were divided into triplets in the format of "Step (Instrument, Action, Anatomy)", specified for each surgical step, similarly as mentioned in [19]. The developed ontology is based on the OntoSPM ontology [20], an emerging common ontology for SPM, which is modeled by making it compatible with a foundational ontology called BFO (Basic Formal Ontology) [21], which is a top-level ontology and provides abstract classes to represent the real-world entities and imported ontologies as shown in Table 1. OntoFox tool was used to extract upper ontological entities [25]. Ontology was built using Protégé (version 5.0.0) [26].

Table 1: Main imported modules of the RAPN ontology

FMA	Foundational Model of Anatomy[22]	Domain of human anatomy. Entities representing anatomy of kidney and surrounding tissue for RAPN workflow.
IAO	Information Artifact Ontology[23]	Information entities in the biomedical domain.
OWL-Time	W3C Time Ontology[24]	Temporal concepts for representing relationships between the surgical entities.

We implemented production rules, with the "IF" and "THEN" statements, to build the reasoning mechanism, which helped to recognize surgical context from the ontology. As shown in Fig. 5, "Step" represents an ontology class, while "hasNextStep", "involvesAnatomicalPart", "hasAction", "hasInstrumentInStep", and "isInPhase" represent semantic relations with "Step". Variables "?x" and "?y" represent real-world instances. Production rules, in total 22, involving RAPN instances, were used for the recognition of the surgical context, e.g. "Actions", "Instruments" and "Phases", based on the unknown step instance ("?x") retrieved from both the prediction of step from the "CRNN" model and the step sequence recognized by the "Sequence" model. As a pre-condition, an ontological relation "hasNextStep" is checked and anatomy was used explicitly specified for each step. The production rules were implemented in Semantic Web Rule Sequence (SWRL) [27].

```
Step(?x) ^ hasNextStep(?x, ?y) ^ involvesAnatomicalPart(?x, Anatomy) →
hasAction(?x, Action) ^ hasInstrumentInStep(?x, Instrument) ^ isInPhase
(?x, Phase)
```

Fig. 5: An exemplary SWRL rule format

"Deep-Onto" network was implemented in Keras (version 2.0.2) with TensorFlow backend (version 1.3.0) [28], OWL (version 3.5.0) [29] and Pellet (version 2.3.3) [30] API to perform reasoning on the ontology.

2.4 Data acquisition, video annotations and data preparation

The videos on RAPN were acquired with the da Vinci Xi surgical system (Intuitive Surgical Inc., CA, USA) at European Institute of Oncology (IEO, Milan, Italy) from September 2016 to June 2017. We recorded 9 videos of RAPN, at 24 FPS with a length of 82.49 $\pm$ 37.54 minutes and the 578x720-pixel HD quality from the da Vinci Xi endoscope with a camera of 8 mm size and 30° angle. The procedures were performed by 4 senior Urologists ("ODC", "GM", "VM", "DB").

Recorded videos are annotated using Anvil annotation tool [31] with workflow entities i.e. the "Ontology class" as shown in Table 2. Each track specifies different surgical workflow entities, in synchrony. Videos are annotated frame-by-frame representing each entity as an individual instance, e.g. "mobilization", as a controlled vocabulary for the ontological class "Step". The definition of workflow entities for annotations was obtained from a journal article [18] and suggestions from the expert Urologist ("RB"), who annotated the videos.

To develop "Nephrec9" dataset[1], first, we split the 9 full RAPN videos into small videos of 30 seconds or 720 frames, processed at 24 FPS. We extracted

[1] available at https://doi.org/10.5281/zenodo.1066831

a total of 1262 videos (996,373 frames). We manually removed 254,800 frames with heavy motion blur, e.g. quick change in the camera position, specular reflections for instruments and tissue surfaces, and frames occluded with heavy smoke. As shown in Table 3, we developed two dataset, "D1" and "D2". "D2" was used to exploit inter-subject variability.

Table 2: Ontology class definitions and video annotations

Ontology class	Definitions
Phase	Major objectives to accomplish the procedure as per standard procedural workflow e.g. after "Tumor Resection" phase, where the tumor is removed, "Renorraphy" phase is performed, which consists of suturing the remaining tissues.
Step	Steps are required to complete phases of the surgical procedure. Each step consists of a specific action, anatomy, and instrument at a specific instance. For example, during "Tumor Exposure" phase, the surgeon makes the "marking" (Step) of the "kidney capsule" (Anatomy) by "marking" (Actions) through the "fenestrated bipolar" (Instrument-Left).
Instrument	Instruments are annotated based on its usage during a step of the surgery and its appearance in surgical videos. We considered robotic instruments, Left and Right robot arm e.g. "fenestrated bipolar" and "monopolar curved scissors". We also considered "laparoscopic Bulldog", which comprises many frames of the recorded videos.
Anatomy	Anatomy is annotated based on a surgical step and its appearance in the videos e.g. "resection" (Step) has "tumor" as "Anatomy".
Actions	Actions are annotated based on a surgical step and actions carried out by specific instruments. For example, "cortical suturing is a "Step" performed by the "large Needle Driver" (Instrument) to "suture" (Actions) the "kidney" (Anatomy) during the kidney repair, i.e. "Renorraphy" (Phase).
Assistant-Instrument1 & Instrument2	These annotations represent the usage of laparoscopic instruments, e.g. "aspirator", by assistant surgeons during RAPN

3 Experimental protocols

We performed off-line testing of "Deep-Onto" components, i.e. "CRNN", "Sequence" and "Knowledge" model.

3.1 Experimental protocols for CRNN model

Four experimental protocols have been designed to check the "CRNN" model as shown in Table. 4.

Table 3: Training, validation and test datasets (Number of frames)

No.	Step	Exemplary image	Training set(D1)	Training set(D2)	Validation set(D1)	validation set(D2)	Testing set(D1)	Testing set(D2)
1	Mobilization		42,469	30,268	17,516	14,162	7,237	22,790
2	Dissection		111,560	111,559	47,228	31,710	18,911	34,428
3	Ultrasound		2,243	1,193	2,265	949	262	2,617
4	Marking		20,433	29,304	10,221	4,143	3,870	1,075
5	Clamping		9,743	10,876	4,991	4,334	1,634	1,115
	Unclamping (represented as “clamping”)		2,614	2,613	1,117	1,116	443	443
6	Resection		60,657	74,986	31,194	11,420	9,209	14,652
7	Cortical suturing (represented as “suturing”)		23,422	21,208	10,587	11,464	4,576	5,911
	Midollar suturing (represented as “suturing”)		31,400	38,237	21,155	10,858	5,965	9,423
8	Removal		2,392	1,435	2,833	2,326	493	1,955
9	Reconstruction		13,960	8,968	14,014	12,423	2,564	9,145
10	Drainage		4,080	2,888	3,702	2,081	1,425	4,236

(1) Experimental protocol-1 (EP1): “CRNN” model was trained on 10 steps, as shown in Table 3, out of 14 annotated steps. In this work, two steps i.e. “Identification” step, which is a subsequent step of “Dissection” in the actual annotations, involves indeed dissection of Gerotas Fascia (anatomy) for assessing the tumor location, and “Inspection” step, a subsequent step of “Suturing”, involves checking the implemented sutures, were not considered to

be recognized, since it represents the similar actions, also same anatomical structures without change in the background in the images of the "Dissection" and "Suturing" steps. Due to these reasons, as expected these steps i.e. "Identification" and "Inspection" created more false positives. Considering the similar background, actions and instruments, we also combined the frames of "Clamping" and "Unclamping" steps, and "Midollar suturing" and "Cortical Suturing" steps into two separate classes of "Clamping" and "Suturing" respectively. The "Suturing" class would not have created any impact on the context-awareness since these involves consequent steps involving suturing actions. Conversely, "Clamp" step instructs one to deal with the "laparoscopic bulldog clamp" insertion or removal, which is handled by assistant surgeons. The ground truth information, i.e. steps, in annotations, has been used to verify the predictions of the "CRNN". We measured accuracy and compiled a confusion matrix to obtain prevalence-weighted macro-average (PWMA) precision, recall, and F1 score as evaluation metric for step recognition.
(2) EP2: EP2 was used to recognize ongoing surgical steps, considering the inter-subject variability. We used dataset "D2", where out of 9 videos, videos 1 to 5 used as a training set, 6 and 7 as a validation set, and 8 and 9 as a test set.
(3) EP3: 8-fold LOOCV (Leave one-out cross-validation) was done to check inter-subject variability, 1 video's samples, from "D1", was used as a test set and rest as train set during each iteration of 8 folds.
(4) EP4: EP4 was designed, as "Baseline" experiments, to do the comparison of "CRNN" prediction of the current steps with the fine-tuned CNN-only network i.e. Inception V3.

Table 4: Experimental protocols for "CRNN" model. The "Experimental Protocol (EP)" shows the implemented protocols i.e. "Hold-out 1" (EP1) (Dataset "D1" was used), "Hold-out 2" (EP2) (Dataset "D2" was used) and "LOOCV" (EP3) were used to check inter-subject variability, and "Baseline" (EP4) shows the step recognition with Inception V3.

No.	Experimental Protocol (EP)	Dataset (Total number of frames)			Samples (Total number of extracted feature vectors)		
1.	"Hold out-1" & "D1" dataset	Training set	Validation set	Test set	Training set	Validation set	Test set
		324, 973 (59.26%)	166,823 (30.42%)	56,589 (10.32%)	555	260	70
2.	"Hold out-2" & "D2" dataset	Training set (videos no. 1-5)	Validation set (videos no. 6,7)	Test set (video no. 8, 9)	Training set	Validation set	Test set
		461,498 (62.25%)	172,096 (23.22%)	107,790 (14.53%)	560	172	154
3.	"LOOCV"	-			Leave 1 video's samples out one by one (videos 1-8)		
4.	"Baseline"	Training set	Test set		-		
		633,594 (85%)	107,790 (15%)				

3.2 "Sequence" model

The experiments were carried out to check the prediction of next step based on the predicted current step by "CRNN" during EP1. Furthermore, prediction's accuracy and algorithm execution time were compared with another sequence prediction method [32] i.e. look-up tables.

3.3 "Knowledge" model

To evaluate "Knowledge" model, recognized step sequences and surgical context are verified with the ground truth video annotations by measuring the relative frequency, f_i, as shown in Eq. (2).

$$f_i = \frac{n_i}{N} \tag{2}$$

In Eq. (2), n_i represents the frequency of occurrence of truly recognized surgical context and N represents the total number of actual surgical context presented in video annotation. The surgical context was considered True Positive (TP) if the same context were represented for a specific step in video annotations as ground truth. Otherwise, it was considered as False Positive (FP). We chose 70 samples from the test set, of "D1", to evaluate "Knowledge" model on surgical context recognition i.e. "Instrument", "Phase", and "Actions". "Anatomy" was explicitly grounded in the assertion box (ABOX), so it could be easily retrieved.

4 Results and discussions

We present the results of the individual models of the pipeline which follow the experiment protocols as mentioned in the section 3.

4.1 "CRNN" model

During EP1, the"CRNN" model was tested on the 70 samples of "D1" dataset. As shown in Fig. 6 and Table 5, the pipeline was able to recognize 10 RAPN steps with 0.83 PWMA recall, 0.74 PWMA precision and 0.76 PWMA F1 score and an accuracy of 74.29%. As shown in Fig. 7, discriminative feature of these steps are clustered with minimum relative entropy. "removal" and "ultrasound" steps were not recognized due to the less videos in the test set. "unclamping" step was wrongly recognized as "suturing" and frames with a similar background as "dissection" step. Many frames of the "marking" were wrongly recognized as "dissection" and "suturing", which is confirmed by overlapping clusters in Fig. 7. EP1 shows that frames of the preceding and subsequent steps affect step recognition accuracy due to homogeneous background.

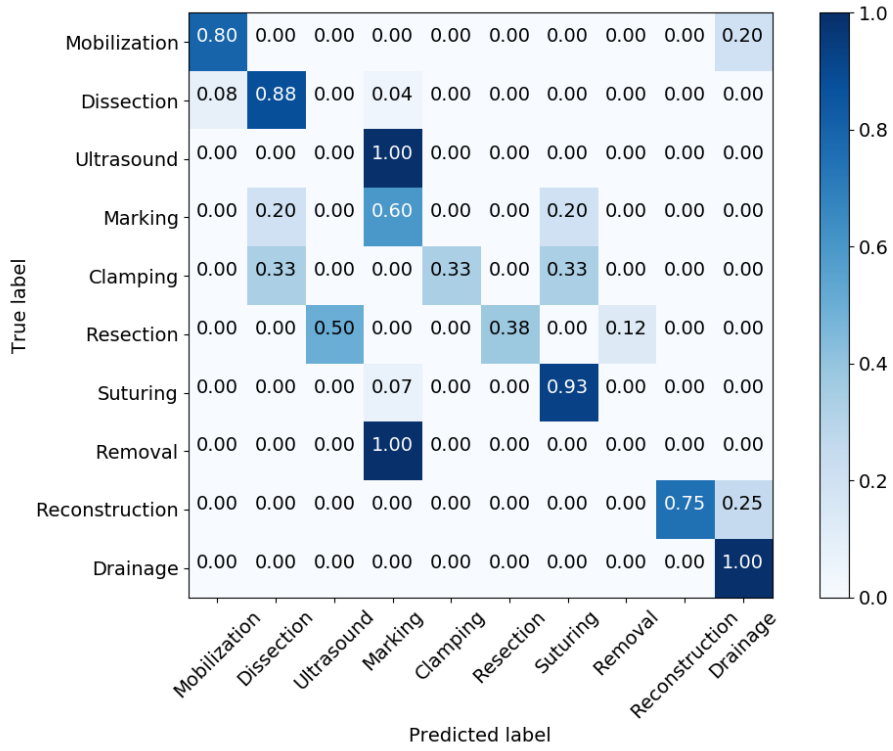


Fig. 6: Normalized confusion matrix (the diagonal values represent individual step recalls for recognition of each step) (EP1)

Table 5: Results on step recognition

Recognized steps	Precision	Recall	F1 score	No. of True Positives	test samples
Mobilization	0.67	0.80	0.73	4	5
Dissection	0.92	0.88	0.88	23	26
Ultrasound	0.00	0.00	0.00	0	1
Marking	0.43	0.60	0.50	3	5
Clamping	1.00	0.33	0.50	1	3
Resection	1.00	0.38	0.55	3	8
Suturing	0.88	0.93	0.90	14	15
Removal	0.00	0.00	0.00	0	1
Reconstruction	1.00	0.75	0.86	3	4
Drainage	0.50	1.00	0.67	2	2

During EP2, the network's recognition accuracy was 67.44% on the validation set and 36.28%±0.1% on the test set. The network was able to recognize 5 steps of the surgery (accuracy in %), "clamping"(100%), "dissection"(83%), "Suturing"(87%), "drainage"(100%) and "ultrasound"(43%). EP2 shows that the dataset is not large enough to learn the variabilities between the subjects. Moreover, the network recognized "ultrasound" step, which have more test samples in "D2". As shown in Table 6, 8-fold LOOCV (EP3) shows the accuracy of 65.58%±6.8%. The cross-validation shows that the network could be able to efficiently recognize inter-subject variability with the increased number of data in the training set. Both the "Hold-out 1" and "LOOCV" results are better than "baseline experiments", which has approximate accuracy of 43.75%±11.2%, that demonstrates CRNN's capacity to provide better recognition of RAPN steps than the CNN-only network.

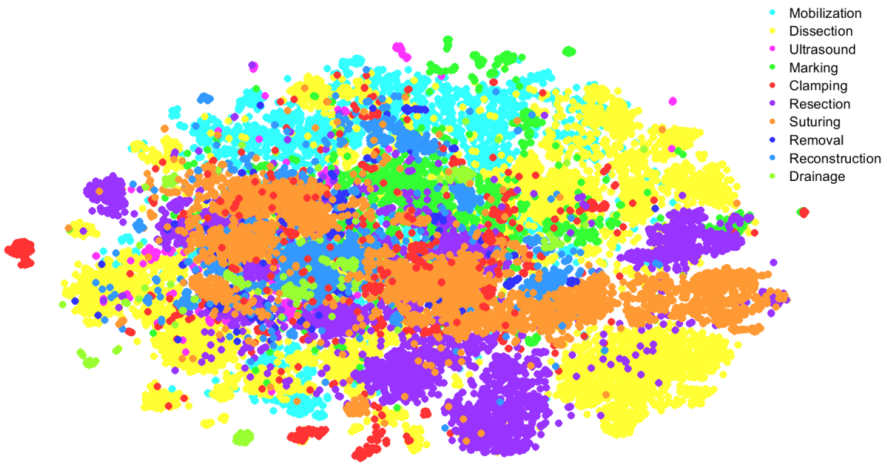


Fig. 7: Extracted feature vectors into two-dimensional space, projected using t-distributed stochastic neighbor embedding (T-SNE) [33] to check the performance of feature extraction. The points were colored according to their true step class labels.

Table 6: RAPN step recognition accuracy scores (in %) for different tests("Nephrec9" dataset)

Experimental protocols	Accuracy (%)
"Hold-out 1"(EP1)	74.28%
"Hold-out 2" (EP2)	36.3±0.1%
"LOOCV" (EP3)	65.6±6.8%
"Baseline" (EP4)	43.8±11.2%

4.2 "Sequence" model

The"Sequence" model and look-up table were 75.7% accurate (true sequence prediction with 53 samples out of 70) in recognizing next steps based on the predicted current steps by the "CRNN" during EP1. "Sequence" model predicted next steps with the larger execution time i.e. 3.41 ± 1.91 seconds as compared to predictions with the look-up table i.e. $1.28 \times e^{-5} \pm 5.43 \times e^{-6}$ seconds. However, considering the choice of the network design as well as further exploitation of the network considering multiple step sequences recognition, and end-to-end training, in this pipeline, RNN is an ideal choice for the sequence prediction task.

4.3 "Knowledge" model

Table 7: Results on surgical context recognition

Recognised context	Relative frequency
Instruments	
Robotic large needle driver	100%
Monopolar curved scissors	50%
Fenestrated bipolar	100%
Laparoscopic bulldog clamp	67%
Actions	
Suture	100%
Dissect	80%
Put	100%
Resect	80%
Clamp	67%
Mark	100%
Phase	
Renorraphy	100%
Hilum dissection	43%
Tumor resection	80%
Tumor exposure	100%
Closure	100%

As shown in Table 7, actions, instruments and phases are recognized with lower relative frequency, i.e. less than 80%, due to wrong recognition of the current step. Steps representing similar anatomy and inverse step relation "hasPreviousStep" was responsible for the classification errors and incorrect recognition of context with the knowledge model.

As shown in Fig. 8, intra-class variations are high, which makes the step recognition task challenging. The large intra-subject variations, which reflects subjectiveness in carrying out surgical steps, affect especially the features extraction process, e.g. a length of frame sequences representing individual

classes are variable. The latter could also be confirmed with the features plotted in Fig. 7, which shows many overlapping clusters. The similar features demonstrate that surgeons do not move the camera much, background textures are similar, and procedures are performed in the narrower region. A large amount of data would also be needed to get the better results. Moreover, understanding of the context including patients, states of devices, anesthesia, team members, etc. could be extended as a future development of an operating room integrated system by including Internet of Things (IoT) approach among all the instruments, room control etc.

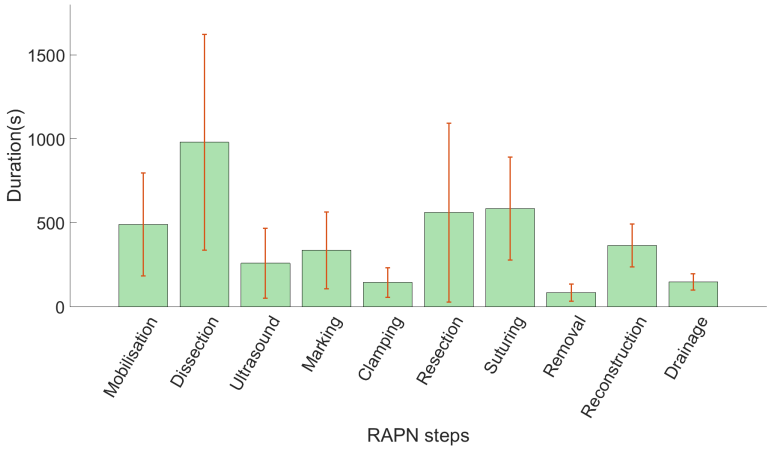


Fig. 8: Intra-class variability of the acquired data, i.e. in 9 videos for 10 RAPN steps, including mean±standard deviation (S.D.), in seconds)

5 Conclusion

We developed a novel "Deep-Onto" network which could allow one to recognize surgical step and its successor step along with the surgical context to some extent, e.g. instruments, actions and phases efficiently. We also developed a new dataset on images of steps of RAPN. We found that the hybrid approach could be useful to do multi-level recognition of the surgical workflow.

Major study limitation was the limited computational memory, i.e. we did not be able to train the network end-to-end with the physical memory of 32 GB. In this work, surgical workflow recognition relies on the correct recognition of ongoing step. However, this study is an essential step towards automatic analysis of surgical workflow. The "Deep-Onto" network is also a modular architecture where other sensor's data, e.g. robot kinematic data could be used more efficiently. The approach could also be extended to other RAS e.g.

robotic cholecystectomy, where the learned network weights could be used for the transfer learning. The ontology could also be extended with the relevant entities of the robotics domain.

As a future work, we will include the temporal information, e.g. optical flows to extract more discriminative features of frame sequences. Moreover, "anatomy" which was currently grounded explicitly in the production rules, if recognized could be used as a pre-condition as a real-time context recognition. As it is hypothesized that a context for the recognition of surgical workflow would be different at each step of the surgery, automatic generation of production rules, e.g. with inductive learning [34], could provide extended capability for adaptive learning on real-world instances.

Conflict of interest The authors declare that they have no conflict of interest.
Ethical standard This article does not contain any studies with human participants or animals performed by any of the authors.
Informed consent Informed consent was obtained from all individual participants included in the study.

Acknowledgements This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No. H2020-ICT-2016-732515. The Titan Xp used for this research was donated by the NVIDIA Corporation.

References

1. Blum T, Feussner H, Navab N (2010) Modeling and segmentation of surgical workflow from laparoscopic video. In: MICCAI International Conference on Medical Image Computing and Computer-Assisted-Intervention, 2010. MICCAI 2010, 13(3):400-7. https://doi.org/10.1007/9783642157110_50.
2. Khurshid AG, Esfahani ET, Raza SJ, Bhat R, Wang K, Hammond Y, Wilding G, Peabody JO, Chowriappa AJ (2015) Cognitive skills assessment during robot-assisted surgery: separating the wheat from the chaff. BJUI, 155(1):166-174. https://doi.org/10.1111/bju.12657.
3. Flin R, Youngson G, Yule S (2007) How do surgeons make intraoperative decisions? Qual Saf Health Care, 16(3):235-239. https://doi.org/10.1136/qshc.2006.020743.
4. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, Forman D, Bray F (2012) Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. European Journal of Cancer, 49(6):1374-403. https://doi.org/10.1016/j.ejca.2012.12.027.
5. Abel EJ, Culp SH, Meissner M, Matin SF, Tamboli P, Wood CG (2010) Identifying the risk of disease progression after surgery for localized renal cell carcinoma. BJU Int., 106(9):1227-83. https://doi.org/10.1111/j.1464-410x.2010.09337.x.
6. Hu JC, Treat E, Filson CP, McLaren I, Xiong S, Stepanian S, Hafez KS, Weizer AZ, Porter J (2014) Technique and outcomes of Robot-assisted Retroperitoneoscopic Partial Nephrectomy: A Multicenter Study. Eur. J Urol. 66:542-549. https://doi.org/10.1016/j.eururo.2014.04.028.
7. Government of Alberta (2018) Robot-assisted partial nephrectomy for renal cell carcinoma : mini review. Available at:https://open.alberta.ca/dataset/0e172257-2820-4eba-9915-f0add1d14f0d/resource/0e537ff8-f84a-4f7f-a00b-ebfbdea0289e/download/ahtdp-partial-nephrectomy-2017.pdf. Accessed 02 May 2018.

8. Lin HC, Shafran I, Murphy TE, Okamura AM, Yuh DD, Hager GD (2005) Automatic detection and segmentation of robot-assisted surgical motions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 8(Pt 1): 802-810. https://doi.org/10.1007/11566465_99.
9. Katić D, Julliard C, Wekerle AL, Kenngott H, Möller-Stich BP, Dillmann R, Speidel S, Jannin P, Gibaud B (2015) LapOntoSPM: an ontology for laparoscopic surgeries and its application to surgical phase recognition. Int J Comput Assist Radiol Surg, 10(9): 1427-34. https://doi.org/10.1007/s1154801512221.
10. Nakawala H, Ferrigno G, De Momi E (2018) Development of an intelligent surgical training system for Thoracentesis. Artificial Intelligence in Medicine, 84: 50-63. https://doi.org/10.1016/j.artmed.2017.10.004.
11. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging. 36(1): 86-97. https://doi.org/10.1109/TMI.2016.2593957.
12. Jin Y, Doi Q, Chen H, Yu L, Qin J, Fu C-W, Heng P-A (2018) SV-RCNet: Workflow recognition from surgical videos using Recurrent Convolutional Network. IEEE Transactions on Medical Imaging 37(5): 1114-1126. https://doi.org/10.1109/TMI.2017.2787657.
13. Cadene R, Robert T, Thome N, Cord M (2016) M2CAI workflow challenge: Convolutional neural network with time smoothing and hidden markov model for video frames classification. arXiv preprint arXiv: 1610.05541.
14. Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2014) Long-term recurrent convolutional networks for visual recognition and description. arXiv:1411.4389. https://doi.org/10.21236/ada623249.
15. Szegedy C, Vanhoucke V, loffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.308.
16. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3): 211-252. https://doi.org/10.1007/s11263-015-0816-y.
17. Kingma DP, Ba J (2017) Adam: A Method for Stochastic Optimization, arXiv:1412.6980.
18. Kaouk JH, Khalifeh A, Hillyer S, Haber G-P, Stein RJ, Autorino R (2012) Robot-assisted Laparoscopic Partial Nephrectomy: Step-by-Step Contemporary Technique and Surgical Outcomes at a Single High-volume Institution. European Urology, 62 (3): 553-561. https://doi.org/10.1016/j.eururo.2012.05.021.
19. Neumuth T, Strau G, Meixensberger J, Lemke H.U., and Burgert O. (2006) Acquisition of process descriptions from surgical interventions In: DEXA 2006 LNCS, vol. 4080, Springer, S., Bressan, J., Kung, R. Wagner (Heidelberg, Germany), 602611.
20. Gibaud B, Forestier G, Feldmann C, Ferrigno G, Gonçalves P, Haidegger T, Julliard C, Katić D, Kenngott H, Maier-Hein L, März K, de Momi E, Nagy D. A, Nakawala H, Neumann J, Neumuth T, Balderrama J. R, Speidel S, Wagner M, Jannin P (2018) Toward a standard ontology of surgical process models. Int J Comput Assist Radiol Surg, 13(9): 1397-1408. https://doi.org/10.1007/s11548-018-1824-5
21. Grenon P, Smith B (2004) SNAP and SPAN: Towards Dynamic Spatial Ontology. Spatial Cognition and Computation, 4(1): 69-104. https://doi.org/10.1207/s15427633scc0401_5.
22. Rosse C, Mejino JLV (2007) The Foundational Model of Anatomy Ontology, in Burger A, Davidson D and Baldock R, Eds. Anatomy Ontologies for Bioinformatics: Principles and Practice, pp. 59-117. https://doi.org/10.1007/978-1-84628-885-2_4
23. Information Artifact Ontology. https://code.google.com/p/informationartifactontology [Online]. Accessed 17 August 2016
24. W3C Time Ontology. https://www.w3.org/TR/owl-time/ [Online]. Accessed on 20.08.2016.
25. Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y (2010) OntoFox: web-based support for ontology reuse. BMC Research notes, 3(1), 175. https://doi.org/10.1186/1756-0500-3-175.
26. Protégé, Stanford Center for Biomedical Informatics Research. Available from: http://protege.stanford.edu. Accessed 12 January 2016.

27. Horrocks I, Patel-Scheider P, Boley H, Tabet S, Grosof B, Dean M (2017) SWRL: A Semantic Web Rule Langauge combining OWL and RuleML. W3C Member Submission 2004, https://www.w3.org/Submission/SWRL. Accessed 23 February 2017.
28. Chollet F (2015) Keras, available at: https://github.com/fchollet/keras. Accessed 05 May 2017.
29. Horridge M, Bechhofer S (2011) The OWL API: A Java API for OWL Ontologies. Semantic Web Journal 2(1): 11-21.
30. Sirin E, Parsia B, Grau BC, Kalyanpur A, Katz Y (2007) Pellet: A pratical OWL-DL reasoner. Web Semantics: Science, Services and Agents on the World Wide Web, 5(2): 51-53. https://doi.org/10.1016/j.websem.2007.03.004.
31. Kipp M (2007) Anvil - A Generic Annotation Tool for Multimodal Dialogue. In: 7th European Conference on Speech Communication and Technology (Eurospeech), pp. 1367-1370.
32. Sun R, Giles CL (2001) Sequence learning: From recognition and prediction to sequential decision making. IEEE Intelligent Systems 16(4): 67-70.
33. Maaten LV, Hinton G (2008) Visualize data using t-sne. J. Mach. Learn. Res. 9:2579-2605.
34. Nakawala N, De Momi E, Pescatori LE, Morelli A, Ferrigno G (2017) Inductive learning of the surgical workflow model through video annotations. In: The IEEE 30th international symposium on computer-based medical systems, CBMS 2017, Thessaloniki, Greece. https://doi.org/10.1109/CBMS.2017.91.