Surgical Phase Recognition: from Instrumented ORs to Hospitals Around the World

Colin Lea, Joon Hyuck Choi, Austin Reiter, and Gregory D. Hager

Department of Computer Science, Johns Hopkins University

Abstract. Automatic segmentation of laparoscopic recordings into sequences of clips is important for analyzing workflow, improving surgical education, and providing surgeons with automated feedback. Despite increasing interest in this problem, current work almost exclusively relies on extensive instrumentation, which is difficult and costly to acquire, and is only evaluated on data collected from individual institutions. In this work we describe two important methodological findings for video-based surgical phase segmentation at-large, and additionally, introduce a new multi-institution surgical phase segmentation dataset. First, we find that a recent Spatiotemporal Convolutional Neural Network (ST-CNN), which captures object motion over short time intervals, is superior at modeling individual surgical phases compared to methods using spatial CNNs or hand-crafted features. Second, we find that a simple Dynamic Time Warping baseline, using the ST-CNN features, outperforms more sophisticated temporal models. We evaluate on the TUM EndoVis dataset, which was collected at a single hospital, and our new EndoTube dataset which was curated from procedures in over a dozen hospitals around the world. While we achieve state of the art performance on EndoVis, we show our approaches do not generalize as well to EndoTube which contains more variability in equipment, video quality, and surgical style.

1 Introduction

Walk into an operating room for laparoscopic surgery and you will see a plethora of devices that can be instrumented and used for automatic workflow recognition. When available, these can be used to recognize surgical events, which may improve operating room efficiency [5], reduce information overload for surgeons [13], or retrospectively analyze surgical workflow [11]. However, most operating rooms do not have these devices or do not have a way of recording the data. In this work we address surgical phase recognition from laparoscopic video which is easy to collect in most ORs. In particular, we focus on offline solutions for large-scale workflow analysis that can be performed across multiple institutions.

Recognizing surgical workflow from video is difficult due to large variability between patients, surgeons, and hospital environments. Patients exhibit substantial variation in appearance due to differences in anatomy such as varying levels of fatty tissue. Surgeons tend to have their own style and may perform surgical phases in different temporal orders. Equipment, such as the endoscope and instruments, may be unique across hospitals, and can result in varying lighting conditions, video quality, and tool appearance. To model these elements of variability, we decompose surgical phase segmentation into two tasks: (1) learn a low-level spatiotemporal model that captures how the environment changes within short time intervals and (2) learn a high-level classifier that captures phase ordering.

Low-level: Individual surgical phases may be ambiguous but are often defined by the configurations of objects (e.g. tools, organs), their spatial relationships, and their motions throughout a sequence. We employ a Spatiotemporal Convolutional Neural Net (ST-CNN) [9], which has shown recent success on other video datasets. This model factorizes video into a spatial component that captures objects in a scene and a local temporal component that captures how these objects change over a short period of time (e.g. 60 seconds). For example, during the *clipping* phase the ST-CNN may capture the applicator tool motion as it applies a clip to the artery. One advantage, compared to spatial-only models (e.g. [15]), is that it explicitly encodes temporal information within the CNN.

High-level: We compare performance with this spatiotemporal CNN in tandem with three classifiers to investigate the importance of high-level temporal information such as sequential phases ordering. First, we compute the most likely surgical phase using per-frame probabilities from the ST-CNN. Second, we employ the constrained Semi-Markov Conditional Random Field from [9] which encodes pairwise relationships between phases. This is similar to the approaches of [15] and [3]. Lastly, we explore a simple Dynamic Time Warping baseline, inspired by Padoy *et al.* [11], which jointly captures low- and high-level temporal information using the ST-CNN activations.

The topic of video-based surgical phase recognition from intraoperative laparoscopic video across multiple institutions has not been studied extensively in the literature but is important for many applications such as retrospective skills assessment [2], workflow analysis [11], and surgical training [8]. Recent datasets, like EndoVis [1] and Cholec80 [15], have been collected from individual hospitals and do not capture the amount of variability seen across institutions. Factors like tool appearance and recording equipment may vary significantly between hospitals and cause video-based models to fail. We introduce a new dataset, EndoTube, which consists of 25 cholecystectomy videos from nine countries and over a dozen hospitals. These videos were carefully curated from procedures uploaded publicly by clinicians. In order to compare with current work, we use the same labels as EndoVis as shown in Figure 1. While performance on this dataset is significantly lower than EndoVis, we highlight the challenges of surgical data captured “in the wild” where there are many more types of variability.

Our primary contributions are: (1) exploring the use of spatiotemporal CNNs for representing surgical phases, (2) comparing three classifiers for capturing high level temporal information, and (3) performing analysis on EndoTube, our new multi-institutional Cholecystetomy dataset.

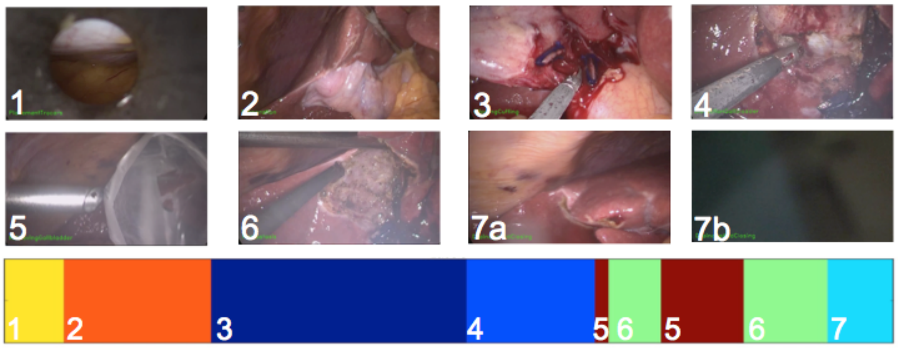

Fig. 1: Example images and sequence labeling from the EndoVis dataset. Phases: (1) Place trocars (2) Prepare Calots triangle (3) Clip/cut cystic artery and duct (4) Dissect Gallbladder (5) Retrieve Gallbladder (6) Hemostasis (7a/7b) Drainage/closure/finish.

2 Prior Work

Despite being a nascent area, there has been substantial recent interest in automated surgical workflow analysis due to publicly available data, promising initial results, and new methods from the computer vision community that may be more capable of modeling complex surgical video.

Recent work by Twinanda *et al.* [15] proposes a CNN-based approach to surgical phase recognition with a Hierarchical Hidden Markov Model. They achieve reasonable performance on the (public) EndoVis and (private) Cholec80 datasets, however, their best results on EndoVis require pre-training on a much larger surgical dataset. While we also use a CNN-based representation, ours explicitly captures temporal information. Dergachyova *et al.* [3] show high performance on EndoVis when combining video and tool data with a Hidden Semi-Markov Model approach, but achieve relatively low accuracy with their video-only variant compared to [15]. Their approach uses hand-crafted image features like color histograms, Histogram of Oriented Gradients (HOG), and Local Binary Patterns (LBP). In both [15] and [3], performance using tool information is relatively low compared to video. Our model achieves superior performance using video, tools, and when modalities are combined.

Earlier work showed that using auxiliary data like tool usage can be effective for workflow analysis [11, 14, 10], however, this requires recording and synchronizing tool data for each surgery which, at scale, is costly and cumbersome.

3 Methods

Our model is comprised of two components: first, we learn a spatiotemporal feature representation using a Convolutional Neural Network that encodes contextual information like tools, organs, and fluids, and models how they change over time. Second, we build a classifier that takes the spatiotemporal features as input and classifies surgical phases.

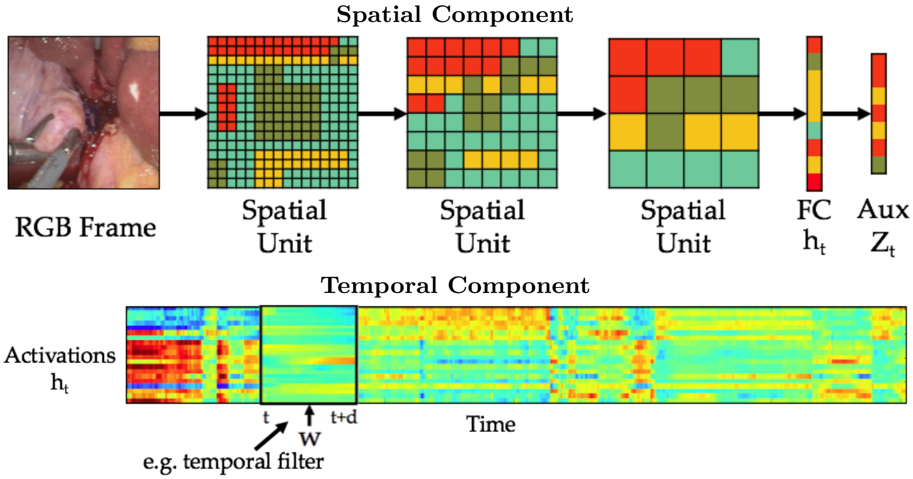

Fig. 2: The Spatiotemporal CNN is factorized into spatial and temporal components. (top) The spatial component consists of spatial units that model the content in each region of an image. (bottom) The temporal component uses the spatial activations, h_t, as input and convolves a set of learned temporal filters. The output is a set of activations, s_t, that encode spatiotemporal information.

While there have been many recent CNN papers, our work is motivated by Lea *et al.* [9] who developed an architecture specifically designed for fine-grained action segmentation. Compared to common CNNs, such as AlexNet [7] and VGG [12], this model better captures spatial relationships when there is a limited amount of data, like in the datasets here. This architecture has significantly fewer parameters which, as a byproduct, makes training and evaluation much faster.

3.1 Spatiotemporal Video Representation

Let I_t be an RGB image for time t from 1 to T, $Z_t \in \{0,1\}^z$ be vector of z auxiliary signals, and $y_t \in \{1,\dots,C\}$ be a phase label. The auxiliary signals can be tool usage information or phase labels as we describe later and the C phase labels are listed in Section 4. Given input image I_t, we compute spatiotemporal activations $s_t \in \mathbb{R}^p$ which is a vector of p latent states.

Spatial Component
The spatial component takes image I_t and outputs intermediate representation $h_t \in \mathbb{R}^f$. This model is composed of spatial units and fully connected layers as shown in Figure 2. The spatial units use convolutional filters to hierarchically model the content in each region of an image. The three spatial units, which were inspired by the VGG architecture [12], consists of a 3×3 convolutional layer, ReLU activation, and 3×3 max pooling. Each colored blocks in the depiction of a spatial unit corresponds to an activation vector in that image region.

The first fully connected layer (FC) consists of latent states, each of which encodes correlations between each region and the corresponding activations in that region. For example, a state may capture the tool being in the top right of the image and the gall bladder being in the middle. Let there be f states in each fully connected layer $h_t \in \mathbb{R}^f$ which correspond to different scene configurations.

The spatial component is trained using auxiliary data Z_t, weight vector $W^{(1)}$ and bias $b^{(1)}$ such that

$$\hat{Z}_t = g_{sp}(W^{(1)} h_t + b^{(1)}) \tag{1}$$

where $\hat{Z}_t$ is the predicted auxiliary signal. When training on the binary tool usage data, $g_{sp}(\cdot)$ is the sigmoid function so that any learned tool can be predicted as *on* or *off*. When training on the phase labels, $g_{sp}(\cdot)$ is the softmax function such that only one class can be chosen at a given time.

Temporal Component

Given the scene activations h_t the temporal component computes a set of temporal activations s_t. We learn temporal convolutional filters that capture how the spatial information changes over time. Empirically, we see these filters capture properties like transitions in state at the start or end of an action.

Each of the l filters, $W_l^{(2)} \in \mathbb{R}^{d \times f}$, is convolved along time with the input, where d is the duration of a filter, $b_l^{(2)}$ is the bias for each filter:

$$s_t^l = \text{ReLU}(\sum_{t'=0}^{d-1} W_{l,t'}^{(2)} * h_{t+t'} + b_l^{(2)}) \tag{2}$$

For notational convenience $h_{t:t+d} \in \mathbb{R}^{d \times f}$ corresponds to the sequence of timesteps from t to $t + d - 1$.

The temporal component is trained using phase labels $Y_t \in \{0, 1\}^c$, where the index of the true class is 1 and all other classes are 0. The output is computed with weights and biases $W^{(3)}$ and $b^{(3)}$:

$$\hat{Y}_t = \text{softmax}(W^{(3)} s_t + b^{(3)}) \tag{3}$$

Predictions $\hat{Y}_t$ correspond to the predicted classes. Note that the input to each classifier is the spatiotemporal activations s_t for all time steps.

Implementation Details

Our network is trained using the cross entropy loss function with ADAM [6], a recent method for stochastic optimization. The three spatial units have $[32, 64, 96]$ convolutional filters, the first fully connected layer has $f = 128$ states, there are $p = 32$ temporal filters, the duration of each filter is $d = 60$ seconds, and the output is $C = 7$ classes. Each input image is $108 \times 108 \times 3$. Parameters are based on those used in [9]. Ideally, the spatial and temporal components could be trained jointly, but due to computational reasons they are learned sequentially. Our model was implemented using Keras[1], a library for developing deep models.

When using multiple data sources, like phase labels and tool information, tools are concatenated at each timestep with the spatiotemporal features s_t. Batch normalization is performed on s_t, using the standard deviation per-feature, in order to normalize the scale of different signals.

[1] Keras: http://keras.io

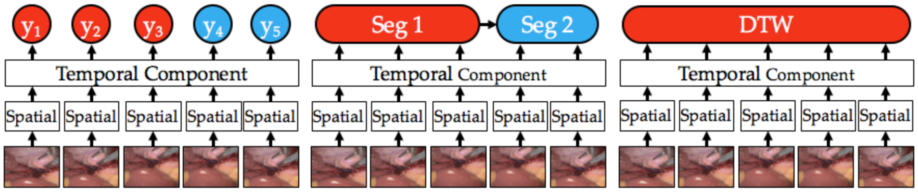

Fig. 3: The full models with the Spatiotemporal CNNs and classifiers. (left) Linear Model (middle) Segmental Model (right) Time-invariant Model

3.2 Surgical Phase Classifier

Our goal is to predict the best phase labeling $\hat{y} = \{\hat{y}_t\}_{t=1}^T$ given spatiotemporal activations $s = \{s_t\}_{t=1}^T$. We compare three classifiers: a frame-wise linear model, the segmental model of Lea *et al.* [9], and a time-invariant model based on Dynamic Time Warping inspired by Padoy *et al.* [11].

1) Linear Model (LM): The output of our ST-CNN, $\hat{Y}_t$, is a set of probabilities corresponding to which phase is active at that time. This model simply takes the most likely phase given the current window of data: $\hat{y}_t = \arg\max_c \hat{Y}_t^c$.

2) Semi-Markov Model (SMM): We use the segmental model of Lea *et al.* [9], a constrained variation on the Semi-Markov Conditional Random Field, which jointly infers the start time s_i, end time e_i, and phase label c_i for each of the M segments in a sequence. Compared to traditional linear chain models like Hidden Markov Models, the Semi-Markov component captures transitions in phase across whole segments. As defined in [9], let K be an upper bound on the number of possible segments in a sequence (e.g. 9 for EndoVis), and P_i be the tuple $\langle s_i, e_i, c_i \rangle$ for the ith segment.

$$\hat{y} = \arg\max_{P_1,\ldots,P_M} \sum_{i=1}^{M} \sum_{t=s_i}^{e_i} \hat{Y}_t^{c_i} \quad s.t. \quad 0 < M \leq K \tag{4}$$

This can be computed very efficiently using the method of [9].

3) Time-invariant Model (DTW): We use an approach inspired by Padoy *et al.* [11] which achieves high performance on surgical phase recognition from tool usage data. While inference in this approach is slower than our other models, because it uses a nearest neighbors approach, it captures both the local changes in spatiotemporal activations within each segment as well higher-level temporal ordering of phases. We use superscript (i) to indicate each trial. For test sequence $s^{(i)}$ we compute the DTW distance [4] to all training sequences $s^{(j)}$:

$$DTW(s^{(i)}, s^{(j)}) = \min_c \sum_{t=1}^{T} \|s_t^{(i)} - s_{c_t}^{(j)}\|_1 \tag{5}$$

where $c = \{c_t\}_{t=1}^T$ are the correspondences between activations in each sequence. Prediction $\hat{y}^{(i)}$ is computed by propagating the labels from sequence j that has the smallest DTW distance such that $\hat{y}_t^{(i)} = y_{c_t}^{(j)}$ for all times t.

4 Datasets

EndoVis Dataset
The EndoVis surgical phase recognition dataset [1], from the Technical University of Munich (TUM), consists of video, tool usage, and phase labels for seven laparoscopic cholecystectomy procedures. The procedures were performed by a small set of surgeons at the same hospital and have similar workflow. Figure 1 shows an example of each phase. In six of seven procedures the phase order is: *Place Trocars*, *Prep*, *Clip/Cut*, *Dissect*, *Retrieval*, *Hemostasis*, *Retrieval*, *Hemostasis*, *Drainage/finish*. In one video, there is only one instance each of *Retrieval* and *Hemostasis*. This dataset also includes tool usage, which was labeled manually, which indicates the instruments in use at any given time. Tools include: *liver retractor*, *fan retractor*, *alligator forceps*, *PE forceps*, *irrigation rod*, *suction rod*, *scissors*, *retrieval bag*, *plastic clips applicator*, *metal clips applicator*. We evaluate on EndoVis with Leave One Video Out cross validation.

EndoTube Dataset
We introduce, EndoTube, a dataset that addresses the ability of our models to generalize to real-world environments. These videos are curated from full cholecystectomy procedures on Youtube and were labeled using the same phases as EndoVis. All videos include each phase from *Preparing Calots Triangle* through *Retrieval*, but may not include *insert tools* or *finish*. This dataset contains 25 procedures which were performed at 19 hospitals in 9 countries. Some videos are as short as 4 minutes and jump in time between each of the major phases, while others last up to 27 minutes and show the whole surgery. The average video length is 11.4 minutes. We sifted through dozens of videos and selected ones in which none of the core phases are skipped and the edits did not substantially detract from the video. Some videos are intended for surgical training and have extraneous segments such as powerpoint slides at the beginning. These portions are label *null* and are removed after prediction but before computing accuracies.

Data was manually labeled using the phase definitions from EndoVis by one engineer experienced in the surgical domain. The labels were verified by a second engineer who was very familiar with the EndoVis dataset. We perform 5-fold cross-validation such that we train on 20 instances and test on 5.

Metrics
We evaluate using accuracy and segmental boundary distance. Accuracy measures the percentage of a video that is correctly labeled. Twinanda *et al.* [15] proposed boundary distance which measures the percentage of the temporal boundaries that are correctly predicted within a certain interval. The motivation is that temporal phase boundaries are often ambiguous and thus the precise start or end time is not of critical importance. Practically, for each segment,

	Spatial CNN			ST-CNN				
Data source(s)	LM	SMM	DTW	LM	SMM	DTW	[3]	[15]
Video	57.6	78.8	81.2	69.0	77.8	**84.6**	68.1	79.7*
Tools	58.5	76.5	85.7	56.4	78.3	**91.2**	78.9	73.0
Video + Tools	73.7	87.3	92.3	81.8	88.5	**92.8**	88.9	-

EndoVis

	Spatial CNN			ST-CNN		
Data source	LM	SMM	DTW	LM	SMM	DTW
Video	47.9	36.0	63.7	56.3	60.1	62.4

EndoTube

Table 1: Results from (top) EndoVis and (bottom) EndoTube. *[15] achieves 86.0% on EndoVis when pre-training their CNN on a larger dataset and with tool information.

we compute the distance of each true starting time and the closest predicted starting time, and determine if their difference is within a specified threshold. We show results for distance thresholds of $\tau = \{30, 60, 90, 120\}$.

5 Results and Discussion

Table 1 shows our accuracy results on both datasets using the spatial-only and spatiotemporal CNNs. Each row was trained using either video, tool information, or both. Recall, when using video, the auxilliary term Z in the spatial component is the set of phase labels at each timestep. When the true tools are used with the video, the tools are concatenated after the temporal CNN component.

We achieve state of the art results when only using tool data, when combining tool and video data, and when only using video (assuming no pre-training). Our high tool-only results are consistent with the findings of Padoy *et al.* [11] on another Cholecystectomy dataset. Twinanda *et al.* [15] perform better than our results when they train on an unpublished surgical dataset, however, when only training on EndoVis our video-based results are better. Methodologically we see that the Spatiotemporal CNN performs favorably compared to a spatial CNN and the hand-crafted features in [3]. Twinanda *et al.* [15] achieve 56.9% accuracy using AlexNet, 62.6% using a spatial CNN trained on EndoVis, and 65.9% when training on both image and tools. For comparison, our ST-CNN, without a high level temporal model (LM), achieves 69.0%. Furthermore, we see that the DTW-based model achieves notably higher accuracy than the linear or semi-Markov models. DTW captures how the spatiotemporal activations change within each phase which appears to have a large impact on accuracy.

Performance on EndoTube (62.4%) is far lower than EndoVis, but is commensurate with the large increase of variability. We analyzed the results from individual videos and found, on average, the best sequence in each split achieves 90.7% accuracy and worst sequence achieves 33.8%. We achieve worst performance when the video quality is low (e.g. abnormally high contrast) and when

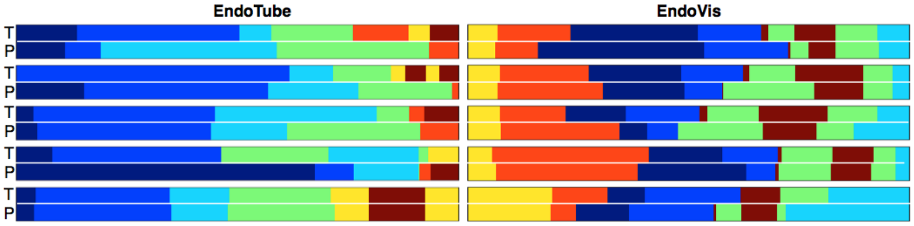

Fig. 4: Example predictions from the EndoTube and EndoVis datasets. The top of each plot depicts the sequence of true phases and the bottom depicts the predicted labels using Dynamic Time Warping. Each color corresponds to a unique surgical phase.

the surgical tools look substantially different than normal. Three of the worst sequences have atypical clipping phases. One surgeon uses thread instead of clips, another uses a unique style of clips, and the third does not use any clips. Despite poor accuracy, we think it is important to include these videos because they address real-world concerns with large-scale workflow analysis.

Table 2 shows the percentage of EndoVis phases that are within τ seconds from the true starting times. We see that most predictions are correct within a reasonable tolerance. When combining video and tool-use, all boundaries are correct within 3 minutes. These videos are 41 minutes on average, so if a phase is correct within 120 seconds, this shift accounts for less than 5% of the video.

Data source(s)	$\leq$ 30	$\leq$ 60	$\leq$ 90	$\leq$ 120	$\leq$ 150	$\leq$ 180
Video	66.2	76.1	82.5	88.8	93.6	93.6
Tools	**90.4**	**90.4**	92.0	93.6	93.6	95.2
Video+Tools	85.2	**90.4**	**92.0**	**95.2**	**98.4**	**100.0**

Table 2: The percentage of predicted label boundaries within the specified distance (in seconds) to the true boundaries on EndoVis using the DTW model.

Figure 4 shows predictions from each dataset. Qualitatively, we see that many errors are simply small temporal shifts. On EndoTube, some predictions (e.g. rows 3 & 5) perform very well whereas others (e.g. rows 1 & 4) perform poorly.

In summary, we make three important observations about surgical workflow analysis. First, despite high performance on single-institution datasets like EndoVis, current models are insufficient for handling the variability on multi-institution datasets like EndoTube. This is a result of an insufficient quantity of data and limitations with the model. Perhaps, new data augmentation techniques could improve performance on these videos. Second, explicitly capturing local temporal information, such as with Spatiotemporal CNN, can improve performance compared to traditional spatial CNNs. Lastly, temporal models like DTW, which jointly capture how our spatiotemporal activations change across time both locally and globally, are beneficial.

References

1. TUM EndoVis. `http://endovissub-workflow.grand-challenge.org/` (2015)
2. Deal, S.B., Lendvay, T.S., Haque, M.I., Brand, T., Comstock, B., Warren, J., Alseidi, A.: Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. The American Journal of Surgery pp. 398–404 (2016)
3. Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., Jannin, P.: Automatic data-driven real-time segmentation and recognition of surgical workflow. International Journal of Computer Assisted Radiology and Surgery (IJCARS) (2016)
4. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: Experimental comparison of representations and distance measures. In Procedings of VLDB Endowment (2008)
5. Franke, S., Meixensberger, J., Neumuth, T.: Intervention time prediction from surgical low-level tasks. Journal of Biomedical Informatics (2013)
6. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference Learning Representations (ICLR) (2014)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems (NIPS) (2012)
8. Kumar, R., Jog, A.S., Malpani, A., Vagvolgyi, B.P., Yuh, D.D., Nguyen, H., Hager, G.D., Chen, C.: Assessing system operation skills in robotic surgery trainees. The International Journal of Medical Robotics and Computer Assisted Surgery. (2012)
9. Lea, C., Reiter, A., Vidal, R., Hager, G.D.: Segmental spatio-temporal cnns for fine-grained action segmentation. European Conference on Computer Vision (ECCV) (2016)
10. Malpani, A., Lea, C., Chen, C.C.G., Hager, G.D.: System events: readily accessible features for surgical phase detection. International Journal of Computer Assisted Radiology and Surgery (IJCARS) (2016)
11. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical Image Analysis (MedIA) (2012), computer Assisted Interventions
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference Learning Representations (ICLR) (2015)
13. Stauder, R., Belagiannis, V., Schwarz, L., Bigdelou, A., Söhngen, E., Ilic, S., Navab, N.: A User-Centered and Workflow-Aware Unified Display for the Operating Room
14. Stauder, R., Okur, A., Peter, L., Schneider, A., Kranzfelder, M., Feussner, H., Navab, N.: Random Forests for Phase Detection in Surgical Workflow Analysis. In: International Conference on Information Processing in Computer-Assisted Interventions (IPCAI) (2014)
15. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. CoRR abs/1602.03012 (2016), `http://arxiv.org/abs/1602.03012`