

Active Learning using Deep Bayesian Networks for Surgical Workflow Analysis

Sebastian Bodenstedt · Dominik Rivoir ·
Alexander Jenke · Martin Wagner ·
Michael Breucha · Beat Müller-Stich ·
Sören Torge Mees · Jürgen Weitz ·
Stefanie Speidel

Received: date / Accepted: date

Abstract *Purpose* For many applications in the field of computer-assisted surgery, such as providing the position of a tumor, specifying the most probable tool required next by the surgeon or determining the remaining duration of surgery, methods for surgical workflow analysis are a prerequisite. Often machine learning based approaches serve as basis for analyzing the surgical workflow. In general, machine learning algorithms, such as convolutional neural networks (CNN), require large amounts of labeled data. While data is often available in abundance, many tasks in surgical workflow analysis need annotations by domain experts, making it difficult to obtain a sufficient amount of annotations.

Methods The aim of using active learning to train a machine learning model is to reduce the annotation effort. Active learning methods determine which unlabeled data points would provide the most information according to some metric, such as prediction uncertainty. Experts will then be asked to only annotate these data points. The model is then retrained with the new data and used to select further data for annotation. Recently, active learning has been applied to CNN by means of Deep Bayesian Networks (DBN). These networks make it possible to assign uncertainties to predictions. In this paper, we present a DBN-based active learning approach adapted for image-based surgical workflow analysis task. Furthermore, by using a recurrent architecture,

S. Bodenstedt · D. Rivoir · A. Jenke · S. Speidel
Department for Translational Surgical Oncology, National Center for Tumor Diseases (NCT), Partner Site Dresden, Dresden, Germany
E-mail: Firstname.Lastname@nct-dresden.de

M. Wagner · B. Müller-Stich
Department of General, Visceral and Transplant Surgery, University of Heidelberg, Heidelberg

M. Breucha · S.T. Mees · J. Weitz
Department of Visceral, Thoracic and Vascular Surgery, Faculty of Medicine and University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany

we extend this network to video-based surgical workflow analysis. To decide which data points should be labeled next, we explore and compare different metrics for expressing uncertainty.

Results We evaluate these approaches and compare different metrics on the Cholec80 dataset by performing instrument presence detection and surgical phase segmentation. Here we are able to show that using a DBN-based active learning approach for selecting what data points to annotate next can significantly outperform a baseline based on randomly selecting data points. In particular, metrics such as entropy and variation ratio perform consistently on the different tasks.

Conclusion We show that using DBN-based active learning strategies makes it possible to selectively annotate data and thereby reducing the required amount of labeled training in surgical workflow related tasks.

Keywords Surgical workflow analysis · Active learning · Bayesian deep learning

1 Introduction

The aim of computer-assisted surgery (CAS) is to provide the surgeon with the right type of assistance at the right moment. For many applications in CAS, such as providing the position of a tumor, specifying the most probable tool required next by the surgeon or determining the remaining duration of surgery, analyzing the surgical workflow is a prerequisite. *Surgical workflow analysis* comprises methods for perceiving and understanding surgical processes in the operating room, generally via data collected from sensors or from human input [15]. Since laparoscopic surgeries are performed using an endoscopic camera, a video stream is always available during surgery, making it the obvious choice as input sensor data for workflow analysis.

Several methods in the state-of-the-art for video-based surgical workflow analysis utilize convolutional neural networks (CNNs) for interpreting the video stream [1, 3, 12, 21, 24]. Deep Neural Networks, such as CNNs, have a high number of parameters that have to be determined during training, which requires a large amount of annotated data. For many tasks in surgical workflow analysis, expert knowledge is often required for labeling data, making it difficult to obtain a sufficient amount of annotations. Motivated by the fact that data without annotations is often readily available, multiple methods for pretraining CNNs using unlabeled data for solving surgical workflow related tasks have been recently proposed [2, 6, 20, 23]. These methods generally exploit information inherent in the unlabeled data to solve an auxiliary task related to the actual problem. Recently crowdsourcing based approaches have been used to successfully create annotations for simple surgical workflow related tasks in laparoscopy, such as tool segmentation [17, 18], locating point correspondences [16] and for assessing skills [5]. More complex tasks, such as surgical phase segmentation, require more task-specific background knowledge, which generally only domain experts, such as surgeons, possess. Often these experts

have limited resources for labeling such data, making it difficult to acquire large, annotated data sets.

A system that could instead actively ask for expert labels only on certain examples, e.g. examples with a high uncertainty, would reduce the total annotation effort and make collecting large, annotated datasets for surgical workflow analysis more feasible. Such a system is called an *active learning* system [4]. During active learning, an initial model is trained using a small amount of labeled data, the *initial training set*. An *acquisition function* then determines through a metric, such as uncertainty, which data points should be labeled next. A new model is then trained on the extended training data [9].

Recently, new methods for estimating uncertainties on the predictions of deep neural networks, such as *Deep Bayesian Networks* (DBN), have been developed [7]. Seeing that such estimates can be used for active learning has motivated Gal et al. [9] to formulate acquisition functions based on DBNs.

In this paper, we investigate if an active learning system based on DBNs can successfully guide the annotation process for image- and video-based surgical workflow related tasks and thereby reduce the number of required labels. For this, we first modify the framework proposed in Gal et al. [9] for laparoscopic instrument presence detection and phase segmentation. Namely, our main contributions are the following:

1. Propose a solution for multi-label annotations with DBN-based active learning
2. Propose a recurrent network for DBN-based active learning with videos
3. Extend the previous network to allow partial annotation of videos
4. Evaluate and compare the proposed methods using the publicly available Cholec80 dataset [21].

To the best of our knowledge, we are the first to apply DBN-based active learning to annotate data related to surgical workflow. Furthermore, as far as we are aware, this is the first work that utilizes DBN-based active learning for video annotation.

2 Methods

In this section, we introduce methods for image-based and video-based active learning for surgical workflow analysis tasks. The basis of our image-based active learning system is a standard CNN that is transformed into a DBN (section 2.1). This serves as basis for performing DBN based active learning on single video frames. To allow active learning on video data, the DBN is further extended into a recurrent DBN (section 2.2). To use the likelihoods of the DBN to select which data points should be labeled next, several different metrics are possible, which are described in section 2.3.

2.1 Bayesian Network

A standard CNN, based on the AlexNet architecture [14] and pretrained on ImageNet (see fig. 1a), serves as a foundation of the proposed system for active learning. We opted to use an AlexNet, as it performed similarly as a ResNet50 during instrument presence detection and phase segmentation, while allowing for faster training. Active learning requires a method for gauging which unlabeled training examples are "difficult" for the current model, e.g. when given an input x , an (softmax) output y and training data D , determining the likelihood $P(y = l|x, D)$ of label l . While neural networks generally do not output a binary class prediction, but instead a fuzzy prediction, e.g. through a sigmoid or a softmax non-linearity, it has been found that these outputs are not suitable as probability estimates [10].

DBNs on the other hand are a mathematically proven concept for estimating likelihoods for predictions [7]. DBNs are deep neural networks with a prior probability distribution, such as a Gaussian prior, placed over the weights W of the network: $P(W)$. The likelihood of a classification is then defined as

$$P(y = l|x, W) = \text{softmax}(f_W(x))$$

where $f_W(x)$ is the output of the network depend on weights W . Inference in DBNs requires the posterior $P(W|D)$, which is extremely difficult to infer. Instead, the posterior can be approximated through Monte Carlo dropout, which is done by performing random dropout on every weight layer during training and testing. Monte Carlo dropout can be shown to be equivalent to performing approximate variational inference, which minimizes the Kullback-Leibler divergence to the true posterior:

$$P(y = l|x, D) = \int P(y = l|x, W)P(W|D)dW \approx \frac{1}{T} \sum_{t=1}^T P(y = l|x, \hat{W}_t)$$

with $\hat{W}_t \sim q_\theta(W)$, where $q_\theta(W)$ is the dropout distribution [7]. In other words, to determining the likelihood of a classification of a sample x during testing, we classify the sample T times using Monte Carlo dropout and average the outputs of the softmax.

The previous CNN that has been extended into a DBN can be seen in fig. 1b. By applying task-specific classification layers to the network, predictions and their likelihood can be estimated.

2.2 Recurrent Bayesian Network

Many tasks in surgical workflow analysis, such as phase segmentation, require that frames are viewed in the context of an entire video or at least in the context of previous frames. Recurrent neural networks (RNN) make such an analysis possible by introducing recurrence into the topology of a network. This allows information from previous frames to contribute to future predictions.

Long short-term memory units (LSTM), a more complex form of the RNN, can learn to strategically remember, but also forget, information from previ-

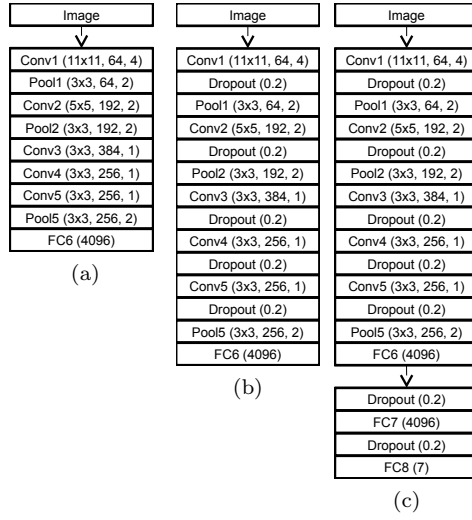


Fig. 1: (a) Standard AlexNet without classification layers, (b) the same AlexNet converted into a DBN through adding dropout layers, and (c) the DBN extended for active learning of laparoscopic instrument presence. For convolutional and pooling layers, the numbers indicate filter size, number of feature maps and step size, for fully connected layers, the number of output units and for dropout layers, the probability of setting a value to 0.

ously seen inputs, while forgoing the problem of vanishing gradients common to RNNs [11]. Combining CNNs with LSTMs makes video-based workflow analysis, by using exclusively deep neural networks, possible [2, 3, 6, 23].

By applying the paradigm described in section 2.1, we can extend the topology of a CNN-LSTM based on AlexNet [14] (see fig. 2a) into that of a Bayesian CNN-LSTM (see. fig. 2b). One approach to perform inference with this network would be to naively apply random dropouts independently to each weight layer for every element in a given sequence. Multiple studies though indicate that such a naive dropout has negative effects on RNNs, such as added noise and a disruption of dynamics [8]. As an alternative, the authors in [8] propose a theoretically grounded variant of dropout for LSTMs. The idea is to sample dropout masks for each layer in the recurrent DBN at the beginning of each sequence and to use the same mask for each time-step (see fig. 3). The naive approach would be equivalent to sampling new masks at every time-step.

This recurrent DBN makes video-based classification possible, while simultaneously allowing likelihood estimations for each classification.

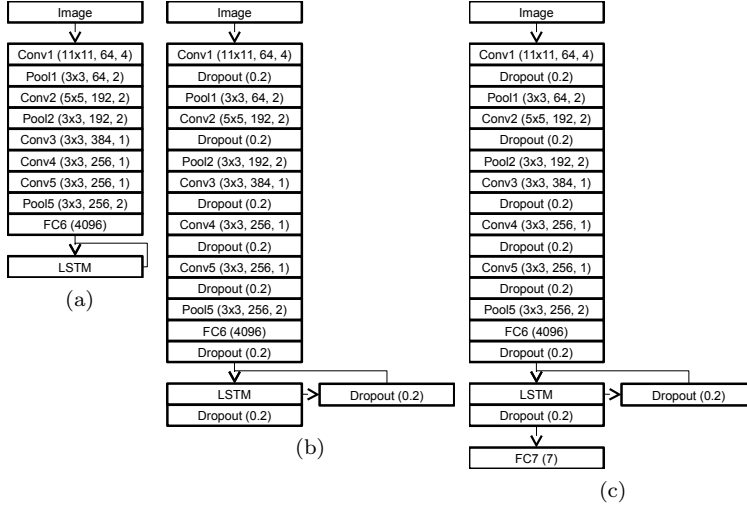


Fig. 2: (a) AlexNet without classifier extended with an LSTM, (b) the same network converted into a recurrent DBN through adding dropout layers, and (c) the DBN extended for surgical phase segmentation. For convolutional and pooling layers, the numbers indicate filter size, number of feature maps and step size, for fully connected layers and LSTMs, the number of output units and for dropout layers the probability of setting a value to 0.

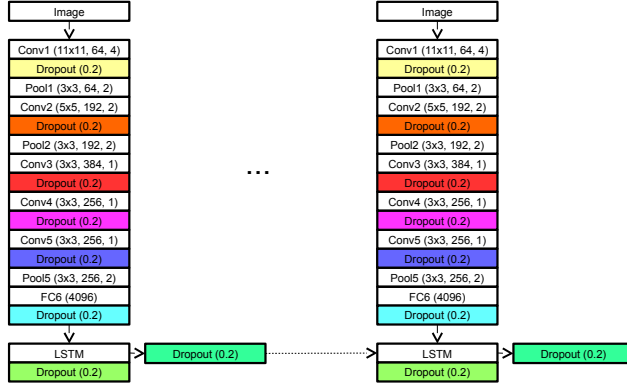


Fig. 3: Modified recurrent DBN in multiple steps of a sequence. The colors indicate identical dropout masks.

2.3 Acquisition Functions

Given a DBN with weights W and a pool with unlabeled data points \bar{D} , the active learning framework uses an acquisition function $a(f_W(x))$, with $x \in \bar{D}$, to determine which data points show high levels of uncertainty. The following

criteria is used to select which data points should be labeled next:

$$x^* = \underset{x \in \bar{D}}{\operatorname{argmax}} a(f_W(x)).$$

The authors in [9] propose multiple acquisition functions that have to be evaluated for their suitability in active learning for surgical workflow tasks.

Variance One simple metric for measuring the uncertainty is to compute the variance of the different likelihoods contributing to the posterior:

$$\operatorname{Var}(x) = \mathbb{E}[(P(y = l|x, \hat{W}_t) - \mu)^2]$$

with $\mu = \mathbb{E}[P(y = l|x, \hat{W}_t)]$. Variance measures how the T likelihood predictions are spread around their arithmetic mean. Here we assume that a large spread corresponds to a large amount of uncertainty.

Variation Ratio (VR) Similarly to variance, the variation ratios also measures the spread of the T predictions, in this case around the mode, i.e. the most common predicted class.

$$\operatorname{VR}(x) = 1 - \frac{f_m}{T}$$

where f_m is the frequency of the mode in the T predictions.

Entropy A further possibility for measuring the uncertainty of the posterior likelihood is using predictive entropy from information theory:

$$\mathbb{H}(x, \bar{D}) = - \sum_l P(y = l|x, \bar{D}) \log P(y = l|x, \bar{D})$$

\mathbb{H} reaches its maximum when the likelihood of all classes becomes equal. Its minimum (zero) is reached when the likelihood of a single class is equal to one.

Mutual Information (MI) An extension of predictive entropy is to examine the mutual information between the posterior and the likelihoods of the T predictions:

$$\mathbb{I}(x) = \mathbb{H}(x, \bar{D}) - \mathbb{E}(H(x, \hat{W}_t))$$

3 Applications

To evaluate the suitability of the DBNs described in the previous section for active learning in workflow analysis tasks, we examine two different applications. In section 3.1 we extended a DBN to perform active learning for laparoscopic instrument presence detection and in section 3.2 a recurrent DBN is used to perform active learning for surgical phase segmentation. For both tasks, the publicly available Cholec80 dataset [21] is used. It consists of 80 videos from laparoscopic cholecystectomies, in which surgical instrument presence and surgical phases have been annotated.

3.1 Instrument Presence

To perform active learning for instrument presence detection, we extend the DBN proposed in section 2.1 with two further fully connected layers (see fig. 1c). The last layer consists out of 7 units, one for each instrument type in Cholec80. Since multiple instruments of different types can be visible in the same video frame, a sigmoid nonlinearity is used on the final layer instead of a softmax. During training, we use a weighted DICE-loss [19] as cost function.

As this is a multi-label problem, when computing the uncertainty of the prediction of a given image using, any one of the acquisition functions $a(f_W(x))$ outlined in section 2.3 will not return a scalar value, but instead a 7-dimensional vector containing the uncertainty of each class. To reduce this vector to a scalar, we examine the suitability of two different methods for aggregation:

$$a_{\text{mean}} = \mathbb{E}(a(f_W(x))) \text{ and } a_{\text{max}} = \max(a(f_W(x))).$$

The frames with the highest uncertainty from \bar{D} are then selected for annotation. a_{mean} would here favor frames with a high certainty across all classes, while a_{max} would favor frames in which one class shows a large amount uncertainty, regardless of the uncertainties of the other classes.

3.2 Phase Segmentation

To allow active learning for phase segmentation, we extend the recurrent DBN proposed in section 2.2 by adding a fully connected output layer (see fig. 2c). The layer has 7 output units, one for each surgical phase in Cholec80, with a softmax nonlinearity. During training, we use cross-entropy as cost function.

We assume that judging the current surgical phase from a single video frame is difficult and prone to ambiguities, we therefore opt to query for annotation for temporally connect segments. For this, we propose two methods for selecting the next queries from \bar{D} .

Video-based A naive approach would be to determine which unlabeled video from \bar{D} has the largest amount of uncertainty and ask an expert to annotate this video completely. Given a video, we compute the uncertainty for each frame and aggregate these uncertainties using either a_{max} or a_{mean} , where a_{mean} would favor videos with a high overall uncertainty and a_{max} would tend to select videos where a single frame exhibits a high uncertainty.

Segment-based Annotating an entire video is a time-consuming process, which is also difficult to parallelize. Instead it would be preferable to select the most uncertain parts of a video and query an expert to just annotate these.

To accomplish this, we divide each video into segments with a length of 5 minutes. During active learning, we determine the uncertainty of each segment in the same manner as described above for an entire video. We then query for the most uncertain segments according to either a_{max} or a_{mean} .

% data annotated	Variance		VR		Entropy		MI		Random
	Max	Mean	Max	Mean	Max	Mean	Max	Mean	
20%	78% (92%)	79% (92%)	79% (92%)	79% (92%)	78% (92%)	79% (92%)	78% (92%)	79% (92%)	78% (92%)
30%	81% (93%)	82% (93%)	80% (92%)	82% (93%)	81% (93%)	82% (93%)	79% (93%)	80% (93%)	79% (92%)
40%	82% (93%)	83% (94%)	81% (93%)	82% (93%)	81% (93%)	82% (94%)	80% (93%)	80% (93%)	81% (93%)
50%	83% (94%)	83% (94%)	82% (93%)	83% (93%)	82% (93%)	82% (94%)	82% (93%)	82% (93%)	82% (93%)
60%	83% (94%)	83% (94%)	82% (93%)	83% (93%)	82% (94%)	82% (94%)	82% (94%)	82% (93%)	82% (93%)

Table 1: Test performance on the instrument presence detection task. Each line shows how the error progresses when more data is selected with one the acquisition functions, in combination with a_{\max} and a_{mean} , or the baseline. The values shown are the weighted F1-score and, in parenthesis, accuracy. The best performances for each line are in bold. The statistical significance is color-coded: Green indicates $p < 0.01$, yellow $p < 0.05$ and red $p \geq 0.05$.

This leads to having incompletely labeled videos during training, which is a problem as the recurrent nature of the network requires that each sequence be trained from the beginning. To account for this, we slightly modify the cost function. Given the output y_i and the correct label l_i of frame i and the cost $C(y_i, l_i)$, we know define $\hat{C}(y_i, l_i) = m_i \cdot C(y_i, l_i)$ with $m_i \in \{0, 1\}$, depending on whether i is annotated or not. This causes frames whose label is unknown at this point, to be excluded from the overall cost, while still preserving their influence on the predictions of the annotated frames.

4 Evaluation

As previously stated, we evaluate our proposed active learning methods for surgical workflow analysis tasks (instrument presence and phase segmentation) on the Cholec80 dataset. For this, we first divide the dataset in 4 subsets of 20 videos each, as outlined in [6]. Each video was sampled at a rate of one frame per second and each frame was downsampled to a resolution of 384×216 pixels. No methods for data augmentation were applied.

During evaluation, we proceed in an identical manner for both instrument presence detection and phase segmentation. We begin by dividing the 4 subsets into a training data set (subsets 1-3) and testing data set (subset 4). We select the first 6 videos (10%) from the training data set and define the remaining 54 video as \bar{D} . The 6 videos and their annotations are used to train an initial DBN using the Adam optimizer [13]. To ensure repeatability and comparability, the layers which have not been pretrained, are initialized with identical values for each experiment. We train for 100 epochs or until the training cost reaches $5 \cdot 10^{-4}$. After training, we note the performance on the test data, namely the weighted F1-score for each class label and accuracy, and proceed to select data points from \bar{D} using one of the acquisition functions in 2.3 and aggregate using either a_{\max} or a_{mean} . New data points are then selected until a further 10% of the training data set has been annotated. We then train the model again from scratch using all the available annotated data, noting the performance on the test data after each training run is completed.

As baseline, we use a fifth acquisition function that selects data points at random. For each task, the baseline is computed 4 times and the results are averaged. Given the averaged baseline and results from one of the introduced acquisition functions, we perform a Wilcoxon signed-rank test to assess statistical significance of the performance changes in the F1-score [22].

Instrument Presence For the instrument presence task, the DBN was trained with a learning rate of 10^{-6} , a L2-norm based weight decay of 10^{-4} and a batch size of 128. Data points in \bar{D} for this task were essentially every frame in the training data, meaning we did not incorporate any knowledge about the structure of the original videos while querying for new frames. We opted to only display results up to 60% as this shows the most drastic differences, most of the methods, including the baseline converged to a F1-score of 83% and an accuracy of 94% shortly after. The results of the active learning process with the different acquisition functions in comparison to the baseline can be found in table 1.

Phase Segmentation For the instrument presence task, the recurrent DBN was trained with a learning rate of $5 \cdot 10^{-5}$, a L2-norm based weight decay of 10^{-4} and a batch size of 128. Similar to the instrument presence task, we opted to display results up to 60% as this shows the most drastic differences, most of the methods, including the baseline converged to a F1-score of 86% and an accuracy of 92% shortly after. The results for the video-based phase segmentation can be seen in table 2a and the results for the segment-based video segmentation in table 2b.

5 Discussion

As tables 1, 2a and 2b clearly show, the DBN-based acquisition functions for active learning generally outperform a baseline based on randomly selecting the next data points. In the case of the instrument presence task, the methods based on a_{mean} seem to outperform their counterpart based on a_{max} , indicating that selecting frames on which the uncertainty is spread among multiple classes is the better strategy. Especially the combination of a_{mean} and the variance-based acquisition function seems to be the method of choice for this task as it consistently achieves the highest performance. Furthermore, it can be observed that actively selecting which data points to include next can lead to a disproportionate increase in occurrence of classes that are underrepresented in the data. Table 3a shows that certain instruments classes, for example the bipolar, scissors and clipper, are selected with a higher frequency when compared to a random baseline. These classes have the lowest rate of occurrence in the dataset.

Similarly, in the phase segmentation task using video-based selection, the methods based on a_{mean} generally outperform their counterpart based on a_{max} . The variance-based acquisition function performs also well on this task, though

% data annotated	Variance		VR		Entropy		MI		Random
	Max	Mean	Max	Mean	Max	Mean	Max	Mean	
20%	66% (77%)	68% (77%)	67% (79%)	68% (76%)	71% (80%)	66% (78%)	67% (81%)	65% (73%)	64% (76%)
30%	68% (78%)	76% (85%)	69% (81%)	73% (82%)	72% (83%)	71% (80%)	75% (84%)	70% (81%)	67% (79%)
40%	73% (80%)	79% (88%)	74% (84%)	79% (87%)	75% (82%)	75% (86%)	76% (83%)	78% (87%)	74% (84%)
50%	77% (87%)	78% (86%)	81% (90%)	82% (90%)	77% (85%)	80% (88%)	77% (85%)	78% (88%)	77% (84%)
60%	80% (87%)	80% (87%)	81% (90%)	82% (91%)	79% (89%)	80% (89%)	80% (88%)	80% (90%)	80% (86%)

(a) Video-based

% data annotated	Variance		VR		Entropy		MI		Random
	Max	Mean	Max	Mean	Max	Mean	Max	Mean	
20%	62% (78%)	59% (75%)	71% (82%)	64% (76%)	68% (79%)	67% (74%)	63% (74%)	71% (78%)	62% (75%)
30%	70% (83%)	68% (79%)	76% (85%)	74% (85%)	79% (87%)	74% (85%)	73% (87%)	73% (85%)	73% (84%)
40%	75% (89%)	75% (87%)	79% (87%)	81% (88%)	80% (88%)	79% (87%)	72% (80%)	79% (88%)	76% (86%)
50%	80% (89%)	77% (86%)	81% (89%)	81% (90%)	83% (89%)	81% (90%)	78% (86%)	81% (88%)	79% (88%)
60%	81% (91%)	84% (91%)	85% (91%)	80% (91%)	83% (90%)	83% (92%)	81% (91%)	84% (92%)	79% (89%)

(b) Segment-based

Table 2: Test performance on the phase segmentation task using completely annotated videos (a) and annotated video segments (b). Each line shows how the error progresses when more data is selected with one of the different acquisition functions, in combination with a_{\max} and a_{mean} , or the baseline. The values shown are the weighted F1-score and accuracy. The best performances for each line are in bold. The statistical significance is color-coded: Green indicates $p < 0.01$, yellow $p < 0.05$ and red $p \geq 0.05$.

the variation rate-based method performs better, actually outperforming the other methods at 50% and 60%.

Interestingly, in the case of the phase segmentation task using segment-based selection, the methods based on a_{\max} seem to be preferable. This indicates that segments containing large peaks of uncertainty seem to add more information than segments with a more distributed uncertainty. Here, the variation ratio and the entropy-based methods seem to perform best. It can also be noted that the segment-based methods generally produces similar results as the video-based methods with less annotated data, meaning that partially annotating videos seems to be a valid strategy. Similarly as during the instrument presence task, it can be observed that certain classes are selected with a disproportionate frequency for annotation. Table 3b shows for example that almost all segments containing frames pertaining to P1 are selected for annotation in the first few rounds. On the other hand, P2 and P4, the longest phases, are selected disproportionately less than with the random baseline.

Overall it can be noted that the acquisition functions based on variation ratio and on entropy, while not always providing the best results, seem to perform consistently well on all tasks, indicating that they might be the best choice when examining a new problem.

6 Conclusion

In this paper, we presented, to the best of our knowledge, the first DBN-based active learning approach for annotating data related to surgical workflow tasks.

% anno- tated	Variance + α_{mean}								Random							
	Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	Bag		Grasper	Bipolar	Hook	Scissors	Clipper	Irrigator	Bag	
10%	60%(9%)	5%(10%)	51%(8%)	3%(15%)	3%(9%)	9%(13%)	5%(8%)		60%(9%)	5%(10%)	51%(8%)	3%(15%)	3%(9%)	9%(13%)	5%(8%)	
20%	57%(20%)	10%(38%)	47%(16%)	3%(32%)	4%(23%)	10%(36%)	7%(23%)		56%(19%)	5%(19%)	54%(18%)	2%(26%)	3%(20%)	7%(23%)	6%(18%)	
30%	56%(30%)	10%(58%)	42%(22%)	4%(68%)	7%(65%)	11%(59%)	8%(39%)		56%(30%)	5%(29%)	54%(28%)	2%(35%)	3%(30%)	6%(32%)	6%(29%)	
40%	55%(40%)	8%(67%)	44%(31%)	4%(79%)	6%(77%)	10%(73%)	7%(49%)		55%(40%)	5%(39%)	55%(39%)	2%(44%)	3%(40%)	6%(41%)	6%(39%)	
50%	54%(49%)	7%(75%)	46%(41%)	3%(85%)	5%(84%)	9%(82%)	7%(55%)		55%(50%)	5%(49%)	55%(49%)	2%(54%)	3%(50%)	6%(51%)	6%(49%)	
60%	53%(58%)	6%(80%)	49%(53%)	3%(89%)	5%(88%)	8%(87%)	6%(60%)		55%(60%)	5%(59%)	55%(59%)	2%(63%)	3%(60%)	6%(60%)	6%(59%)	

(a) Instrument type occurrence

% anno- tated	Entropy + α_{max}								Random							
	P1	P2	P3	P4	P5	P6	P7		P1	P2	P3	P4	P5	P6	P7	
10%	12%(21%)	31%(6%)	8%(9%)	31%(9%)	4%(7%)	11%(12%)	3%(8%)		12%(21%)	31%(6%)	8%(9%)	31%(9%)	4%(7%)	11%(12%)	3%(8%)	
20%	9%(37%)	40%(20%)	8%(20%)	27%(18%)	4%(22%)	9%(24%)	2%(12%)		6%(26%)	37%(18%)	8%(20%)	29%(19%)	4%(21%)	9%(24%)	7%(38%)	
30%	16%(97%)	49%(35%)	5%(20%)	18%(18%)	5%(25%)	7%(27%)	2%(16%)		6%(38%)	41%(30%)	7%(26%)	27%(27%)	4%(26%)	10%(37%)	6%(46%)	
40%	12%(97%)	39%(38%)	7%(37%)	18%(24%)	6%(58%)	13%(65%)	5%(56%)		6%(53%)	42%(40%)	8%(39%)	26%(34%)	4%(35%)	10%(53%)	5%(48%)	
50%	9%(97%)	34%(41%)	8%(51%)	25%(41%)	7%(83%)	12%(77%)	5%(61%)		6%(58%)	42%(50%)	8%(52%)	29%(48%)	3%(40%)	8%(55%)	4%(52%)	
60%	8%(97%)	35%(50%)	9%(74%)	27%(52%)	6%(92%)	11%(84%)	5%(73%)		5%(63%)	40%(58%)	8%(63%)	30%(59%)	4%(60%)	9%(70%)	4%(61%)	

(b) Phase occurrence

Table 3: Changes of occurrence of different classes due to data selection with a DBN-based active learning approach compared to random data selection during the instrument presence task (a) and the segment-based phase segmentation task (a). The values indicate the percentage of samples of each class in the training set before each annotation round and, in parenthesis, the percentage of all occurrence of a class contained in the current training set.

Our focus, in particular, was on instrument presence detection and workflow analysis. Also we presented the first DBN-based active learning approach for video annotation. Furthermore, we showed that our approach for selecting the next data points for annotation outperforms a random baseline and we were able to demonstrate that partially annotating videos is a valid strategy for training CNNs for surgical workflow segmentation.

Even though the results seem promising, we see potential for improving performance. The step size of 10% for selecting data points might not be optimal as it could encourage unnecessary redundancy in the data, as it can be assumed that similar images have a similar uncertainty. Opting for a smaller step size might mitigate this problem. Furthermore, incorporating a form of similarity measure in the acquisition functions might also be appropriate. Currently we retrain our network from scratch after new labeled data has been acquired. An alternative strategy would be to instead fine-tune the existing network, though this requires research into metrics for determining from which previous state to fine-tune.

Conflict of interest S. Bodenstedt, D. Rivoir, A. Jenke, M. Wagner, M. Breucha, B. Müller-Stich, S. Mees, J. Weitz and S. Speidel declare that they have no conflict of interest.

Ethical approval For this type of study formal consent is not required.

Informed consent This article contains patient data from publicly available datasets.

References

1. Aksamentov, I., Twinanda, A.P., Mutter, D., Marescaux, J., Padoy, N.: Deep neural networks predict remaining surgery duration from cholecystectomy videos. In: MICCAI, pp. 586–593. Springer (2017)
2. Bodenstedt, S., Wagner, M., Katić, D., Mietkowski, P., Mayer, B., Kenngott, H., Müller-Stich, B., Dillmann, R., Speidel, S.: Unsupervised temporal context learning

- using convolutional neural networks for laparoscopic workflow analysis. arXiv preprint arXiv:1702.03684 (2017)
3. Chen, W., Feng, J., Lu, J., Zhou, J.: Endo3d: Online workflow analysis for endoscopic surgeries based on 3d cnn and lstm. In: *Computer Assisted Robotic Endoscopy*, pp. 97–107. Springer (2018)
 4. Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. *JAIR* **4**, 129–145 (1996)
 5. 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* **211**(2), 398–404 (2016)
 6. Funke, I., Jenke, A., Mees, S.T., Weitz, J., Speidel, S., Bodenstedt, S.: Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. In: *First International Workshop, OR 2.0*, p. 85. Springer (2018)
 7. Gal, Y.: *Uncertainty in deep learning*. University of Cambridge (2016)
 8. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. In: *NIPS*, pp. 1019–1027 (2016)
 9. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: *ICML* (2017)
 10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. *ICML* (2017)
 11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
 12. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE TMI* **37**(5), 1114–1126 (2018)
 13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
 14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc. (2012)
 15. Lalys, F., Jannin, P.: Surgical process modelling: a review. *IJCARS* **9**(3), 495–511 (2014)
 16. Maier-Hein, L., Kondermann, D., Roß, T., Mersmann, S., Heim, E., Bodenstedt, S., et al.: Crowdtruth validation: a new paradigm for validating algorithms that rely on image correspondences. *IJCARS* **10**(8), 1201–1212 (2015)
 17. Maier-Hein, L., Mersmann, S., Kondermann, D., Bodenstedt, S., Sanchez, A., Stock, C., Kenngott, H.G., Eisenmann, M., Speidel, S.: Can masses of non-experts train highly accurate image classifiers? In: *MICCAI*, pp. 438–445. Springer (2014)
 18. Maier-Hein, L., Ross, T., Gröhl, J., Glocker, B., Bodenstedt, S., et al.: Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In: *MICCAI*, pp. 616–623. Springer (2016)
 19. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *3D Vision (3DV), 2016 Fourth International Conference on*, pp. 565–571. IEEE (2016)
 20. Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *IJCARS* pp. 1–9 (2018)
 21. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE TMI* **36**(1), 86–97 (2017)
 22. Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics Bulletin* **1**(6), 80–83 (1945). URL <http://www.jstor.org/stable/3001968>
 23. Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of cnn-lstm networks. arXiv preprint arXiv:1805.08569 (2018)
 24. Zisimopoulos, O., Flouty, E., Luengo, I., Giataganas, P., Nehme, J., Chow, A., Stoyanov, D.: Deepphase: Surgical phase recognition in cataracts videos. In: *MICCAI*, pp. 265–272. Springer (2018)