Multimed Tools Appl (2018) 77:8061–8079
DOI 10.1007/s11042-017-4699-5

Learning laparoscopic video shot classification for gynecological surgery

Stefan Petscharnig[1] · **Klaus Schöffmann**[1]

Received: 11 October 2016 / Revised: 6 April 2017 / Accepted: 7 April 2017 /
Published online: 22 April 2017

Abstract Videos of endoscopic surgery are used for education of medical experts, analysis in medical research, and documentation for everyday clinical life. Hand-crafted image descriptors lack the capabilities of a semantic classification of surgical actions and video shots of anatomical structures. In this work, we investigate how well single-frame convolutional neural networks (CNN) for semantic shot classification in gynecologic surgery work. Together with medical experts, we manually annotate hours of raw endoscopic gynecologic surgery videos showing endometriosis treatment and myoma resection of over 100 patients. The cleaned ground truth dataset comprises 9 h of annotated video material (from 111 different recordings). We use the well-known CNN architectures *AlexNet* and *GoogLeNet* and train these architectures for both, surgical actions and anatomy, from scratch. Furthermore, we extract high-level features from AlexNet with weights from a pre-trained model from the Caffe model zoo and feed them to an SVM classifier. Our evaluation shows that we reach an average recall of *.697* and *.515* for classification of anatomical structures and surgical actions respectively using off-the-shelf CNN features. Using GoogLeNet, we achieve a mean recall of *.782* and *.617* for classification of anatomical structures and surgical actions respectively. With AlexNet the achieved recall is *.615* for anatomical structures and *.469* for surgical action classification respectively. The main conclusion of our work is that advances in general image classification methods transfer to the domain of endoscopic surgery videos in gynecology. This is relevant as this domain is different from natural images, e.g. it is distinguished by smoke, reflections, or a limited amount of colors.

Keywords Video classification · Deep learning · Convolutional Neural Network

✉ Stefan Petscharnig
stefan.petscharnig@itec.aau.at

Klaus Schöffmann
ks@itec.aau.at

[1] Alpen-Adria Universitat Klagenfurt Fakultät fur Technische Wissenschaften, Universitätsstraße 65-67, Klagenfurt, 9020, Austria

1 Introduction

In recent years, endoscopic surgery procedures as well as imaging technology have advanced rapidly. These advances enable physicians to perform minimally invasive surgeries. As a side-effect, the recoded surgery videos benefit the surgeons' work, as they provide a great basis for documentation, training of young surgeons, and medical research. Prior work supporting these aims has been conducted by our research group in the sector of endoscopic video analysis, such as a subjective quality assessment for the impact of compression on the perceived semantic quality [13], instrument classification in laparoscopic videos [17], or extraction and linking of endoscopic key-frames to videos [3, 23]. In this work, we restrict ourselves to a very specific field in minimally invasive surgery in the context of gynecology. In particular, we base our work on videos showing surgical treatment of myoma resection and endometriosis. Our aim is to lay a baseline for (semi-) automatic documentation for aforementioned surgical interventions. Therefore, we want to achieve semantic classification of video shots displaying surgical tasks and various anatomical structures relevant to gynecological surgery. Standard hand-crafted features lack the expressive power for use cases of high-level classification in this domain [2]. On the contrary, CNNs have been successfully used for such problems in general image and video domains [7, 25]. Multiple models have been proposed for semantic classification of video shots, i.e. single frame, early fusion, late fusion, and slow fusion [6]. The importance of deep learning in medical image analysis and content-based processing and analysis of endoscopic images and video also is apparent from the work of Litjens et al. [9] and Muenzer et al. [12] respectively.

As stated above, we aim at creating a baseline for semi-automatic documentation and therefore restrict ourselves to a single-frame model. Hence, the driving question behind our research is:

> *How well do CNN-based single-frame models for semantic shot classification in the field of gynecological surgery, a special domain of laparoscopic surgery, perform?*

In order to answer the aforementioned question, we identify frequent surgical tasks and anatomical classes in cooperation with medical experts from the regional hospital (LKH) Villach in Austria. Based on this expert knowledge and over 100 video recordings of surgical treatments, we generate a data set with scenes of surgical actions and anatomical structures in gynecological surgery. The data set comprises 13 different semantic classes (five anatomy and eight action classes) and consists of about 9 h of annotated video material. Furthermore, we base our work on two well-known CNN architectures: AlexNet [7] and GoogLeNet [25]. For both subsets, surgical action and anatomy, we adapt the classification layer of the aforementioned networks, train the networks from scratch, and evaluate the predictive performance of the resulting networks. The division of action and anatomical structures is reasonable, as we employ a single label prediction model and surgical actions almost always show anatomical structures. We also evaluate the usage of high-level CNN features (from AlexNet classification as well as fully connected layers *fc6* and *fc7*) for a multi-class SVM classifier in the domain of endoscopic surgery videos in gynecology.

This work is novel, as there is no comparison of different CNN models and SVM classifiers using CNN-extracted features for the use case of shot classification in gynecologic surgery. We expect that advances in the general domain transfer to our specialized use case, in particular we think that GoogLeNet achieves a better predictive performance than AlexNet. Furthermore, we expect that the off-the-shelf CNN features do not work as good for classification as the CNN models do. Another contribution of this work is a detailed discussion of important semantic content classes in the expert-domain of minimally invasive gynecologic surgery. This is relevant to colleagues working in the field of medical video analysis. The remainder of this paper is structured as follows. First, we discuss related work in medical imaging on the topics of computer-aided diagnosis, transfer learning, and semantic video classification. In Section 3, we describe the data annotation process as well as the data used for training and testing the CNN models and SVM. Details for learning are presented in Section 4. We evaluate the results in Section 5 and draw conclusions and outline possible future work in Section 6.

2 Related work

For the use case of classifying interstitial lung diseases, Li et al. [8] provide a simple CNN model containing a single convolutional layer. They yield per–class precision and recall between 0.8 and 0.9 for classification into five classes (normal, emphysema, ground glass, fibrosis, and micro-nodules) outperforming the SIFT feature as well as Restricted Boltzmann Machines. Anthimopoulos et al. [2] propose a deep CNN model containing five convolutional layers for the classification of CT images into seven classes of interstitial lung diseases (healthy, ground glass opacity, micronodules, consolidation, reticulation, and honeycombing). Their results imply that, for this use case, their CNN approach outperforms other CNNs as well as state-of-the-art methods using handcrafted features. In the work of Yan et al. [29], a multi–stage deep learning framework is presented. Using the proposed framework, the authors try to solve the problem of body-part recognition in MRI images. In total, they achieve best performance regarding recall, precision and f–score compared against logistic regression, SVMs, and CNNs. The importance of CNNs in medical applications is also apparent from their use within other applications such as nucleus segmentation [28], polyp detection in colonoscopy videos [15], microcalcification detection in digital breast tomosynthesis [22], mitosis detection in breast cancer histology [1], and short–term breast cancer risk prediction [19]. Our work is delimited to the aforementioned research as in contrast to the classification of a state (e.g., healthy or consolidation, type of tissue), we aim at classifying both, anatomical structures and surgical actions. Furthermore, there haven't been any efforts made regarding the classification of images extracted from laparoscopic surgery videos. Fine tuning and transfer learning effects of CNNs are covered in recent literature by Shin et al. [24] as well as Tajbakhsh et al. [26]. These pieces of work are based on the use cases of lymph node detection, interstitial lung disease classification, polyp detection and image quality assessment in colonoscopy, pulmonary embolism detection in computed tomography images, and intima-media boundary segmentation in ultrasonographic images. Their results imply that CNNs are suitable for computer aided diagnosis problems, and transfer learning from large-scale annotated natural image datasets is beneficial for performance (which according to our preliminary studies does not apply to the problem of scene classification). For colonic polyp classification, Riberio et al. [21] proposed transfer learning using off-the-shelf CNN features. Based on high-level CNN features (from CNNs trained for object recognition), Ng et al. [4] use semantic fisher vectors for semantic classification of natural video scenes. Their results reach state-of-the-art performance on MIT Indoor and SUN datasets. For a large-scale YouTube video dataset, Karpathy et al. [6] give an overview on scene classification models based on CNNs, i.e. single frame, late fusion, early fusion and slow fusion. Their results imply that the naive single frame model (which is agnostic to temporal information)—despite it simplicity—already provides a strong performance. Ng et al. [30] compare single frame models for scene classification with slow fusion and LSTM-based models. In the domain of cataract surgery videos, Quellec et al. [20] propose a temporal segmentation and recognition of tasks. The temporal segmentation is based on the detection of idle phases, which is achieved by nearest neighbor search in a reference dataset. Primus et al. [11] provide a video segmentation for endoscopic surgeries based on analysis of spatial and temporal motion changes. For the use case of cholecystectomy, a special form of laparoscopic surgeries, Primus et al. [18] provide a rule-based method to temporally segment a surgery into different phases. The recognition of number and kind of used instruments (which is topic of their previous work [17]) act as main indication for a surgery phase. Shot boundary detection in cholecystectomy surgery videos using Gaussian Mixture Models and a Variational Bayesian Algorithm is investigated by Loukas et al. [10]. The work of Twinanda et al. [27] also focuses on the use case cholecystectomy. They successfully apply CNNs, SVMs and HHMMs for detection of surgical phases. The envisioned classification is different from the use cases mentioned above, as in cholecystectomy there are predefined surgical phases, whereas in other fields of laparoscopic surgery (such as as gynaecology) there is no general consensus for such surgical phases. Moreover, we do not aim at defining shot boundaries. We provide the work most related to this by ourselves [16] in which we already have preformed an exploratory investigation of shot classification in the laparoscopic surgery domain. However, we did no distinction between surgical actions and anatomical structures which resulted in poor performances in the anatomical structure classes.

3 Laparoscopic gynecology video database

For this work, we analyze 111 different gynecological surgery videos. These videos contain scenes of laparoscopic endometriosis treatment and laparoscopic myoma resection and have a duration in the range of 20 min to 6 h. Analysis and discussion with medical experts for gynecology at the regional hospital (LKH) Villach (Austria) have resulted in the identification of two main aspects for the individual scenes: action and anatomy.

Anatomy This type of video scene features little or almost no surgical actions apart from moving tissue and organs. Purpose of diagnosis scenes is the assessment of pathologies on specific organs, such as ovaries, uterus, or liver. Hence, diagnosis scenes are relevant for documentation purposes of the disease as well as its treatment. These scenes are important for medical research and teaching purposes. A second aspect of diagnosis scenes is to document the treatment outcome, i.e. which actions are performed, or how the tissue after treatment looks like. Additionally to disease treatment documentation, diagnosis scenes can be valuable whenever postoperative complications occur. According to our use case of myoma resection and endometriosis treatment, we identify the following (sub-) classes as diagnosis scenes of interest: *Uterus*, *Ovaries*, *Oviduct*, *Liver* and *Colon*. Please note that this list of classes is no comprehensive list of anatomical structures visible in the surgery videos, but it covers the most important organs which are encountered during surgical treatment. For an overview on anatomical structure classes, please refer to Fig. 1.

Action The class of surgical action video scenes feature significant interaction with the patient's tissue and organs using a variety of different surgical instruments. These scenes represent the main physical work for the surgeon. Their automatic classification is relevant for documentation and even more for teaching purposes of certain operation techniques.

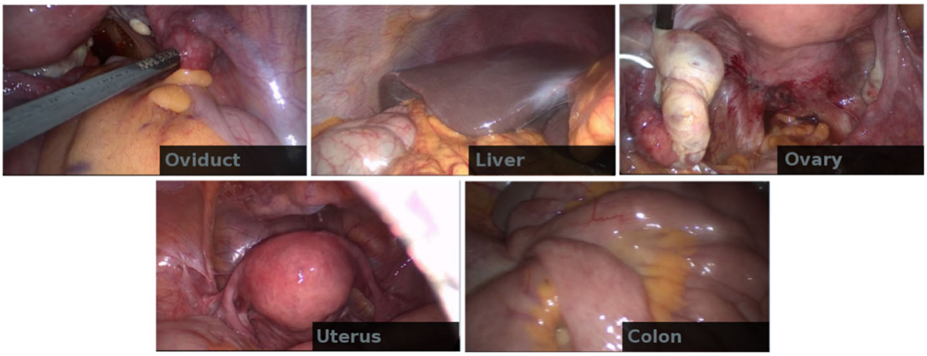


Fig. 1 An overview on anatomical structure classes uterus, ovary, oviduct, liver, and colon. The frames are extracted from the annotated data set

The main aspect of these scenes is the use of medical instruments, e.g. suction & irrigation device, graspers, monopolar needles, needleholders, or scissors. We identify the (sub-) classes *Suction & Irrigation*, *Suture*, *Dissection (blunt)*, *Cutting*, *Cutting (cold)*, *Sling*, *Coagulation*, and *Injection* as the most common surgical actions during laparoscopic endometriosis treatment and myoma resection in our dataset (see Fig. 2). Of course, there are several other actions to be performed, such as tissue extraction, or stapling, but as mentioned before, we are interested in the most common and most important actions.

3.1 Annotation process

We derive the best matching class for a single shot implicitly by camera position and the current action, e.g., the action in the center of the image or the organ which is inspected

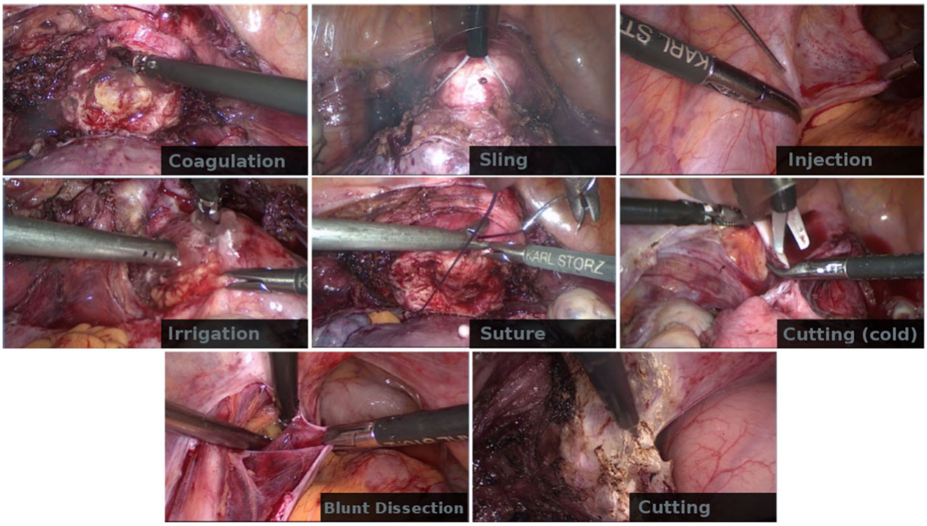


Fig. 2 An overview on surgical actions coagulation, sling, injection, irrigation, suture, cutting (cold), cutting, and blunt dissection. The frames are extracted from the annotated data set

by a surgeon is the action or object of interest. With the surgical action classes, there is the issue that a shot is likely to contain frames that could be classified as a diagnosis class as well. For example, suturing the ovary may contain images with the ovary without a surgical needle, or the suture is not clearly visible. On the one hand, this frame does not look like it belongs to a suturing shot, but on the other hand it indeed does belong to the suturing shot as the image has been recorded in its context. For the annotation of our dataset, we choose to stick to the latter case and annotate such frames as the surgical task by defining begin and end of the surgical action. Each frame from beginning until the end of a shot is labeled with the corresponding shot label for the class it belongs to. Due to this circumstance, the dataset also may contain blurry frames or frames in which instruments may cover huge parts of the camera. We argue that these frames are nonetheless part of the corresponding shot and thus correctly labeled. Prior to the annotation process, our annotators have been trained by medical experts. The annotations are cross-validated by a single annotator and trimmed in length or corrected when necessary. We do not filter blurry or irrelevant frames, as we are interested in a baseline evaluation without any preprocessing (except for resizing and center cropping) of the raw video frames. Thus, we leave the temporal dependencies within the annotated scenes intact.

3.2 Semantic content classes

Due to legal restrictions, we are not able to publish the used dataset. In order to allow for partial repeatability, we give a detailed explanation of the individual classes in the following.

Suction & Irrigation. These scenes feature the use of the suction and irrigation tube. Irrigation has the purpose to clean tissue in order to provide a clean field of view for the surgeon. Main visual feature is a ray of liquid. The suction action is quite the opposite to irrigation. It is used to absorb liquids. Classification problems in this class arise, whenever the suction and irrigation tube is used for positioning tissue or palpation.

Suture. The main characteristic of suturing scenes is the visible surgical needle and the suture. In general, the surgical needle can be of round or straight physical shape. During the process of suturing, the surgical needle often is only partially visible, if at all. The suture can vary in type, thickness, and color. An additional characteristic of these scenes is the use of the knot pusher, which is preceded by a scene where suture and low motion is visible.

Cutting (cold). Scenes of cold cutting, as annotated in this dataset, feature the separation of tissue with a sharp instrument, such as a scalpel or a scissor. Characteristic to this type of scenes is the use of multiple instruments: the instrument used for dissection itself (e.g. scissors) and grasper for fixation of tissue. This characterization applies to cutting and blunt dissection as well.

Cutting. Cutting scenes show surgical separation of tissue by using electro-surgery technology such as mono-polar needles. Occasionally, a bright dot can be seen at the top of the instrument. A low to medium emission of smoke emerges from coagulated and separated tissue.

Dissection (blunt). Blunt dissection scenes feature the use of blunt instruments for the dissection of tissue. In our dataset, no specific tools can be bound to this action – the surgeon uses two or more blunt tools.

Sling. This class contains scenes of separation of the uterus for extraction. The electrical sling itself has an insulation which may look just like a special kind of suture. The coarse procedure of this surgical action is (i) introduction of the sling, (ii) positioning around the cervix, and eventually, (iii) thermal dissection. The thermal dissection features a significant amount of smoke. After this dissection, coagulation and suturing are required in general.

Coagulation. These type of scenes show coagulation by electro-surgical surgery methods. These scenes feature medium to high emission of smoke. The used instruments for this action do vary. For example, surgeons can use graspers or scissors which implies an additional difficulty for the classification of such scenes.

Injection. These scenes feature the injection of liquid into the patient's tissue in order to minimize traumata. The injection needle is visible as thin straight piece of shiny rounded metal. The tissue around the tip of the needle typically inflates after the injection.

Uterus. The uterus is the main organ of interest during myoma resection. In endometriosis treatment, the uterus can also be of interest in the adenomyosis disease pattern. The videos sequences of the class uterus feature an inspection of the uterus.

Ovary and *Oviduct* These classes are again of diagnostic nature. They feature image frames of clearly visible ovary. They are especially important for endometriosis disease and diagnosis of adhesions.

Liver and *Colon*. These two organs also are inspected during endometriosis diagnosis and treatment.

Out of 111 raw gynecological surgery videos, we manually annotated 1,105 shots consisting of 822,918 different video frames resulting in about 9 h of annotated video scenes. As already mentioned, the annotators have been trained by medical experts and the annotated scenes have been checked partly by the experts. Tables 1 and 2 give an overview on the annotated medical video database including class ID, class name, and short semantic description for each action and anatomy class. Moreover, they contain information about the distribution of annotations on a per-class basis, i.e. number of annotated shots, number of annotated frames, average scene duration, and standard deviation. Most frequent actions observed in this dataset are *Suction and Irrigation*, *Coagulation*, and *Cutting (Cold)*. *Suture* is the leading class in terms of annotated video duration. On average, *suturing* scenes have longest duration, scenes of *Cutting (Cold)* are the shortest. The variance within the individual classes arises from surgery circumstances, such as intervention complications, or patient anatomy. Due to the high variance of video sequence length (class–wise compared to average duration), no statistically significant conclusions can be drawn from the individual scene length.

4 Frame-based shot classification

For this work, focus on the feasibility of endoscopic shot classification of laparoscopic surgery videos in gynecology with CNNs. Moreover, we investigate how end-to-end trained CNN with a problem-specific classification output layer perform against off-the-shelf CNN features.

Therefore, we use a single-frame scene classification model allowing us to investigate the influence of different network architectures and the quality of extracted high-level CNN features for the application of SVMs. We base our shot classification on two different network architectures: AlexNet [7] and GoogLeNet [25], which are designed for general purpose image classification and trained for the 1,000 classes of the ILSVRC dataset. AlexNet features input image patch sizes of 227×227 pixel. It consists of five convolutional layers, MAX pooling, local response normalization, dropout and three fully connected layers. The

Table 1 An overview on the annotated dataset with surgical actions: class id, class name, number of shots, number of frames, average duration in seconds, standard deviation of duration in seconds, and class description

ID	Class	Shots	Frames	t_{avg} [s]	t_{sd} [s]	Description
1	Dissection (blunt)	58	35,517	24.49	32.30	Blunt dissection of tissue (e.g by tearing it apart)
2	Coagulation	212	84,786	16.00	16.09	Application of coagulation in order to close a wound
3	Cutting (cold)	271	26,388	3.89	4.32	Dissect tissue with a sharp instrument (e.g. scissors)
4	Cutting	106	92,653	34.96	49.96	Thermally dissect tissue (e.g. with monopular electrodes)
5	Hysterectomy (Sling)	25	68,466	109.55	71.27	Dissection of large parts of tissue with an electrical sling
6	Injection	52	52,355	40.27	26.66	Injection with a needle
7	Suction & Irrigation	173	73,977	17.10	24.63	Application of the suction and irrigation tube
8	Suture	92	321,851	139.94	77.51	Process of suturing

last fully connected layer is task-specific. Thus, for our experiments, the number of output neurons is altered to 5 and 8 output neurons for anatomy and action models respectively. Apart from this, the remaining network structure remained unaltered. The GoogLeNet architecture features inception modules with dimensionality reduction. In total, there are 22 parametrized layers and five pooling layers. Below the stacked inception modules (each reducing the image resolution) there is a convolutional low-level feature extraction expecting input patches of 224×224 pixels. The end of the network features a fully connected network. Analogous to the procedure with AlexNet, the network architecture remains unchanged except for the adaptation of the classification layer.

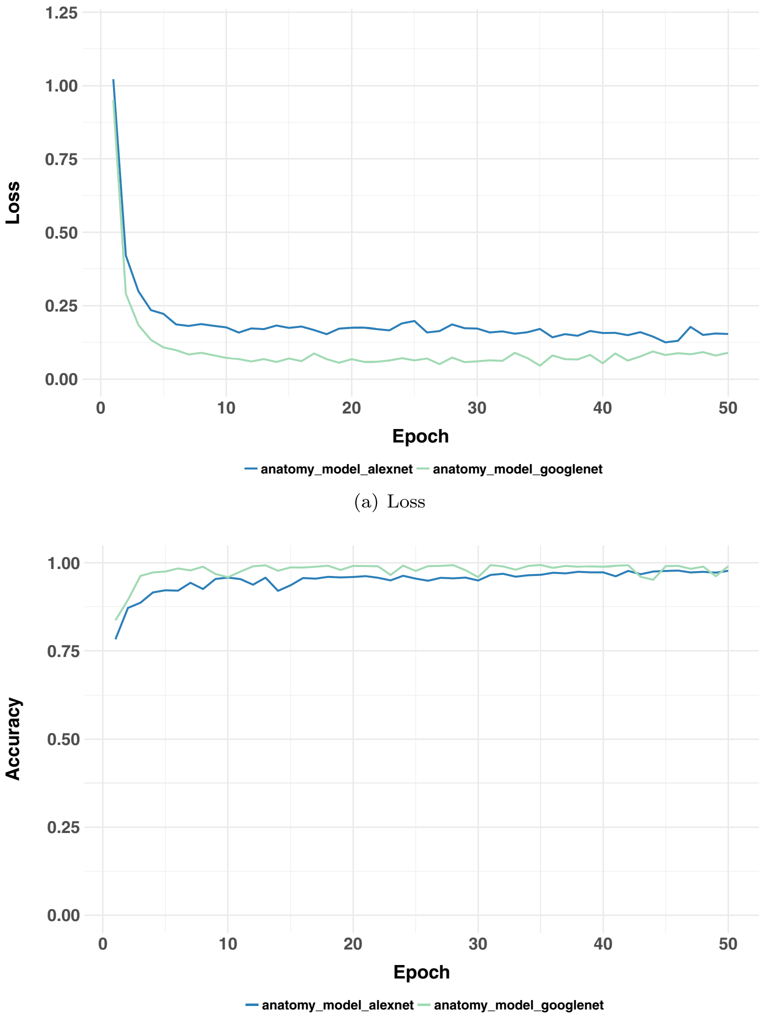
We prepare the video database for training and evaluation, which simply means that we extracted a square center crop of each video frame and then resized it to 256×256 pixel. Thus, we save computational resources for resizing and cropping at training time. We furthermore split the endoscopic video dataset into a test and a training set for each, anatomy and action images. For the split, we considered the test set to contain approximately 10% of the annotations. To ensure a diverse test set, we set a minimum number of images per class.

Table 2 An overview on the annotated dataset with anatomical structures: class id, class name, number of shots, number of frames, average duration in seconds, standard deviation of duration in seconds, and class description

ID	Class	Shots	Frames	t_{avg} [s]	t_{sd} [s]	Description
1	Colon	6	7,285	48.57	56.60	Clearly visible colon
2	Liver	10	3,378	13.51	12.39	Clearly visible liver
3	Ovary	52	28,460	21.89	25.15	Clearly visible ovary
4	Oviduct	8	4,797	23.99	29.78	Clearly visible oviduct
5	Uterus	40	23,005	23.01	41.80	Clearly visible uterus

For the anatomy subset this means that we included at least 500 unique frames per class in the test set and for action, we included at least 5,000 unique frames. The anatomy test set thus comprises 6,874 unique frames, the action test set comprises 57,205 unique frames. The remaining video frames are used to generate the test set. Please note that (as apparent from Tables 1 and 2) for both, action and anatomy subsets, the distribution of number of scenes and frames is highly imbalanced. For example, the action *Suture* is a frequent action and features long scene durations. We thus feature a high number of suturing frames in the database. On the other hand, there are actions such as *Blunt Dissection* featuring a very small number of unique frames. For the test set, this imbalanced distribution perfectly models our use case, as the frequently occurring classes are tested more thoroughly. For the training set, we eradicate this imbalance by a combination of undersampling (dropping frames randomly from the training set) and naive oversampling (duplicating frames randomly). To create the training set, we choose the number of training examples per class to 100,000 images for the action subset and 10,000 images for the anatomy subset. We define that classes containing more unique images than the training set size per class are overrepresented classes. Otherwise a class is underrepresented. For overrepresented classes, we (uniformly) randomly choose the corresponding number of images from the remaining images without returning the chosen images to the set we chose from. The data loss is negligible as we are dropping many near-duplicate images. For the underrepresented classes, we choose images with returning them to the set we chose from (uniformly) at random. We ensure that each annotated image is included in this process by pre-filling the training set with one image of each underrepresented class. This process resulted in 50,000 training images (generated from 33,732 unique images) for the anatomy model and 800,000 training images for the action model (generated from 486,771 unique images).

For implementation of the machine learning approaches (CNN and SVM), we use Caffe [5] and OpenCV [14]. At training time, we feed the network image patches of its expected size (224 pixel squares for GoogLeNet, 227 pixel squares for AlexNet). These image patches are crops chosen at random from the training images featuring a size of 256×256 pixels. As additional data augmentation, we also use Caffe's mirror feature at training time. For optimization, we use the *Adam* optimization method with initial learning rate of 0.001 and momentum parameters 0.9 and .999 Other hyperparameters like weight decay are not altered from their respective values as shipped with the AlexNet and GoogLeNet models. The training is performed on a machine featuring an Intel(R) Core(TM) i7-5960X CPU 3.00GHz processor, 64GB of DDR-4 RAM, a Samsung SSD 850 pro and a NVIDIA GeForce GTX TITAN X graphics card. For AlexNet, we use a batch size of 100 images per batch. For GoogLeNet the batch size is set to 50 images per batch. For both, AlexNet and GoogLeNet, we train action and anatomy models from scratch. This system takes approximately ten days for training of all models and SVMs. The training loss and validation performance of the CNNs is depicted in Fig. 3 for the anatomy models and in Fig. 4 for the action models. The x-axis shows the training epoch. The y-axis shows loss and accuracy respectively. At each epoch, we measure average loss of the epoch and validation performance. For the anatomy models, the loss and accuracy curves bottom out after approximately 10 epochs. In the surgical action models, the training loss for the GoogLeNet network rises after 2 epochs. Longer training of AlexNet has the same effect. Also the accuracy of the model drops with higher numbers of epoch. We think this behavior origins in overfitting. For anatomical structures, this is less a problem as the individual classes are less diverse. We select the models for evaluation with respect to least train loss and highest training accuracy.

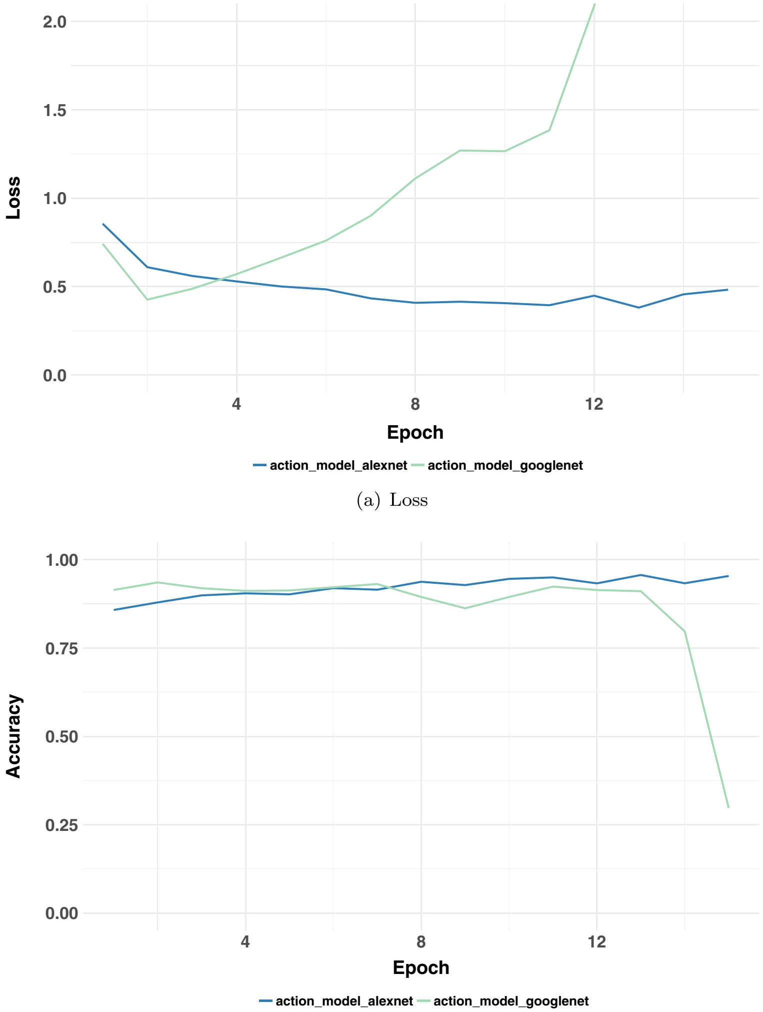


(a) Loss

(b) Accuarcy

Fig. 3 Loss and accuracy for anatomy models based on AlexNet and GoogLeNet CNN architectures for 50 epochs

For the SVM learning process, we classifiy the training set with the AlexNet model with our weights and with off-the shelf weights which have been pre-trained for ImageNet classification. We extract feature vectors from three different locations of the network: the *vector of class probabilites*, the layer *fc7*, and the layer *fc6* as input for SVM training and testing. For simplicity we refer to these vectors as *class*, *fc7*, and *fc6* respectively. We use OpenCV's C_SVC, which enables n-class classification with penalty multiplier for outliers. We do not set specific weights per class, thus we are treating misclassification of each class equally. This approach is reasonable, as we use a balanced training set. We use a *linear* SVM kernel, as this kernel worked best within preliminary studies. As termination criterion, we set the maximum number of iterations to 1,000 and the tolerance to 10^{-6}.



(a) Loss

(b) Accuarcy

Fig. 4 Loss and accuracy for action models based on AlexNet and GoogLeNet CNN architectures for the first 15 epochs

5 Evaluation

For evaluation, we use the trained models of AlexNet and GoogLeNet architectures for action and anatomy classification as well as SVM classifiers trained on high-level CNN feature vectors *fc6*, *fc7*, and *class* from the AlexNet architecture. As weights, we use off-the-shelf weights that ære trained for ImageNet classification. In order to compare the predictive performance of the networks and the SVM approach, we use class-based precision and recall as well as average precision and average recall values over all classes. Evaluating precision and recall in a class-based manner has the advantage that the imbalance of the classes in the test set is taken into account. For the calculation of precision, recall, and f-value of class i, we determine TP_i (true positive classification of class i), FP_i (number of false positive predictions for class i), and FN_i (number of false negative predictions of class i). We also calculate the probability that the true class is among the top three predictions. We refer to this probability as Recall@3, which we can not evaluate for the SVM approach as the OpenCV interface does not allow for that.

For the evaluation, we create an own validation set consisting of approximately 70,000 frames by choosing five representational scenes per class. Please note that these scenes are neither in the training nor in the test set. Thus, this additional set validates the generalization capabilities of the approaches. The validation set size for action and anatomy is 50,988 and 21,568 images respectively. For a class distribution within the validation set, please refer to Table 3.

For a detailed and class-based performance overview, please consult Table 4 for the surgical action classification and Table 5 for anatomical structure classification.

On average, GoogLeNet achieves the best results for **surgical action classification** in terms of Recall, Precision, f-value and Precision@3. However, there are classes where other approaches work better. For example, AlexNet is better at the classification of *Coagulation*. We think that origins in the fact that tissue after coagulation and cutting with a monopolar needle device looks very similar and is distinguished by the used instruments only (which are not visible on each frame in the scenes and also appear frequently in other scenes). GoogLeNet interprets these instruments more likely to be contained in other scenes than AlexNet. The SVM approach using layer *fc6* is better at classes *Injection* as well as *Suction & Irrigation*. These two classes are special, as they feature most reflections. We think that features from AlexNet trained on the ILSVRC dataset better map reflections as the models trained on a database where reflections occur constantly.

For **anatomical structure classification**, GoogLeNet also dominates the average performance in terms of Recall, Precision, f-value, and Precision@3. Interstingly, if we look at Recall@3, AlexNet slightly surpasses GoogLeNet at *Colon*, *Ovaries*, and *Uterus* classes. The other two classes, *Oviduct* and *Liver* are dominated by GoogLeNet. Considering the small number of anatomical structure classes, Recall@3 is not that expressive for the anatomy subset when the distances are that small as we observe them in the cases GoogLeNet performs worse than AlexNet. In terms of f-value, the combination of precision and recall, GoogleNet dominates in all but the *Liver* class, where the SVM approach using *fc7* features dominates with a value of .909 compared to .879. The same approach yields

Table 3 Overview on the validation data set

Class ID	Action class	#imgs	Anatomy Class	#imgs
1	Blunt Dissection	1,620	Colon	1,396
2	Coagulation	2,037	Liver	1,846
3	Cutting Cold	655	Ovaries	3,174
4	Cutting	2,634	Oviduct	3,032
5	Hysterectomy (Sling)	5,119	Uterus	1,336
6	Injection	4,446		
7	Suction & Irrigation	1,475		
8	Suture	7,508		

Each class consists out of five scenes

Table 4 Detailed evaluation results for the action subset

	1	2	3	4	5	6	7	8	Avg.
Recall									
AlexNet	.577	**.431**	.176	.858	.621	.177	.323	.590	.469
GoogLeNet	**.792**	.322	**.354**	**.962**	**.923**	.406	.484	**.690**	**.617**
SVM Class	.002	.000	.116	.073	.399	.000	**.532**	.003	.141
SVM fc7	.631	.308	.180	.687	.626	.549	.236	.299	.440
SVM fc6	.632	.302	.272	.612	.743	**.682**	.406	.470	.515
Precision									
AlexNet	.447	**.295**	.229	.607	.768	.741	.135	.593	.477
GoogLeNet	.566	.260	**.246**	**.838**	**.860**	**.881**	.254	**.812**	**.590**
SVM Class	**.571**	.000	.018	.492	.405	.500	.075	.458	.315
SVM fc7	.470	.207	.110	.568	.463	.681	.215	.574	.411
SVM fc6	.517	.290	.150	.552	.681	.651	**.292**	.707	.480
f-value									
AlexNet	.504	**.350**	.199	.711	.687	.285	.190	.591	.440
GoogLeNet	**.660**	.288	**.290**	**.896**	**.891**	.555	.333	**.746**	**.852**
SVM Class	.005	.000	.031	.128	.402	.000	.131	.006	.088
SVM fc7	.539	.248	.136	.622	.532	.608	.225	.393	.413
SVM fc6	.569	.296	.193	.581	.711	**.666**	**.340**	.565	.490
Recall@3									
AlexNet	.820	.732	.214	.966	.882	.392	.818	.895	.715
GoogLeNet	**.956**	**.857**	**.647**	**1.00**	**.972**	**.762**	**.988**	**.920**	**.888**

For class IDs of the action classes, please refer to Table 1. Bold numbers indicate the top performance within a class

good performance regarding recall for the class *Uterus*. With a value of .874, the features of *fc6* layer also provide a good precision for *Oviduct* classification.

Our results further imply that introduction of an additional SVM classifier does not improve prediction results on average when introducing more sophisticated neural networks. This off-the-shelf feature approach looses performance in terms of recall per class and mean precision compared to GoogLeNet CNN. Interestingly, for actions, the more basic layer *fc6* works better than the more abstract features *fc7* and *class* achieving very poor performances. For anatomical structures, the layer *fc7* works best out of the three evaluated features which are used as SVM input. We observe that the GoogLeNet architecture is superior to the AlexNet architecture and SVM Classifiers.

Hence, this gives a strong indication that improvements of CNN methods in the general domain of image classification lead to improvements in the specialized domain of laparoscopic surgery image classification. Also, off-the-shelf features from AlexNet and linear SVMs slightly outperform AlexNet training from scratch when the right layer is chosen. We think this originates in the training set. This set is correctly annotated, but not fully noise-free considering individual images. Comparing surgical action to anatomical structure classification performance, it is obvious that anatomical structures perform much better in overall performance. We think this originates in the very complex nature of surgical action scenes compared to more static scenes featuring anatomical structures and the agnostic of the temporal dimension.

Table 5 Detailed evaluation results for the anatomy subset

	1	2	3	4	5	Avg.
Recall						
AlexNet	.652	.596	.858	.442	.528	.615
GoogLeNet	**.795**	.862	**.888**	**.623**	.743	**.782**
SVM Class	.554	.601	.484	.562	.581	.556
SVM fc7	.663	**.891**	.755	.374	**.801**	.697
SVM fc6	.572	.854	.712	.412	.697	.649
Precision						
AlexNet	.595	.765	.546	.800	.613	.664
GoogLeNet	**.805**	.896	**.747**	.839	**.619**	**.781**
SVM Class	.461	.659	.591	.860	.273	.569
SVM fc7	.792	**.927**	.561	.862	.475	.724
SVM fc6	.751	.882	.535	**.874**	.408	.690
f-value						
AlexNet	.622	.670	.667	.569	.568	.619
GoogLeNet	**.800**	.879	**.811**	**.715**	**.676**	**.776**
SVM Class	.503	.629	.532	.680	.372	.543
SVM fc7	.722	**.909**	.644	.522	.596	.679
SVM fc6	.649	.868	.611	.560	.514	.641
Recall@3						
AlexNet	**.979**	.694	**.986**	.773	**.989**	.884
GoogLeNet	.977	**.928**	.965	**.868**	.981	**.944**

For class IDs of the action classes, please refer to Table 2. Bold numbers indicate top performance within a class

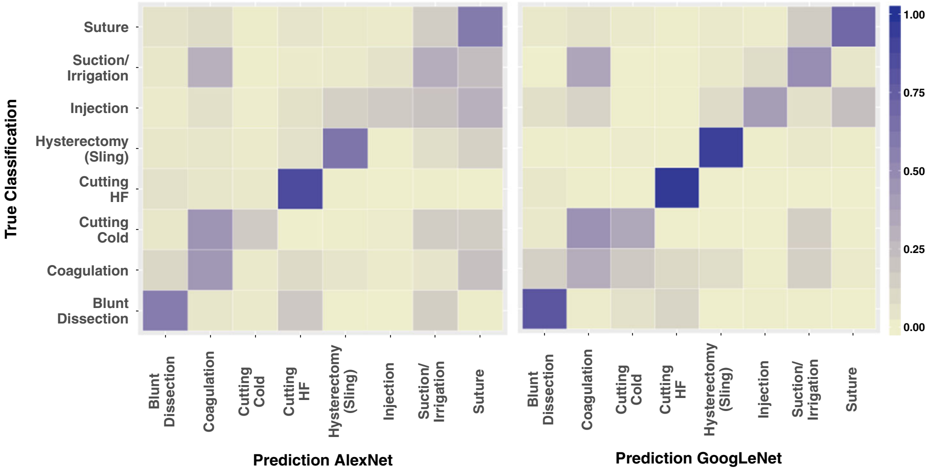


Fig. 5 Confusion matrices for action models based on AlexNet and GoogLeNet CNN architectures

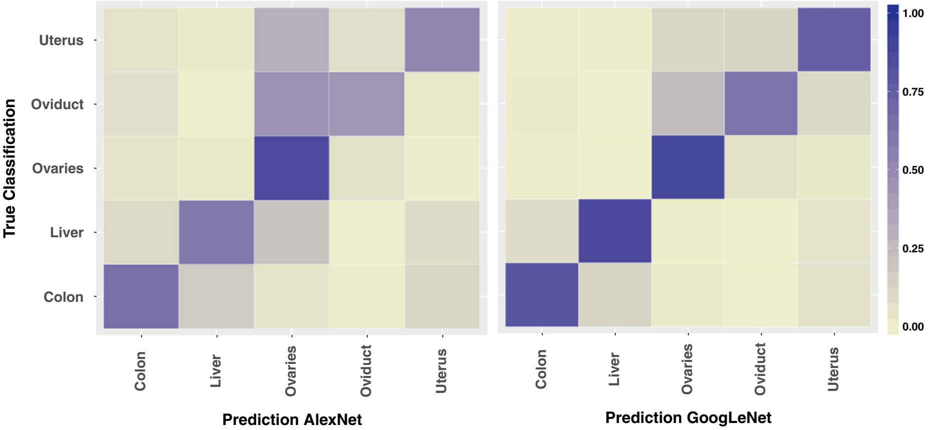


Fig. 6 Confusion matrices for anatomy models based on AlexNet and GoogLeNet CNN architectures

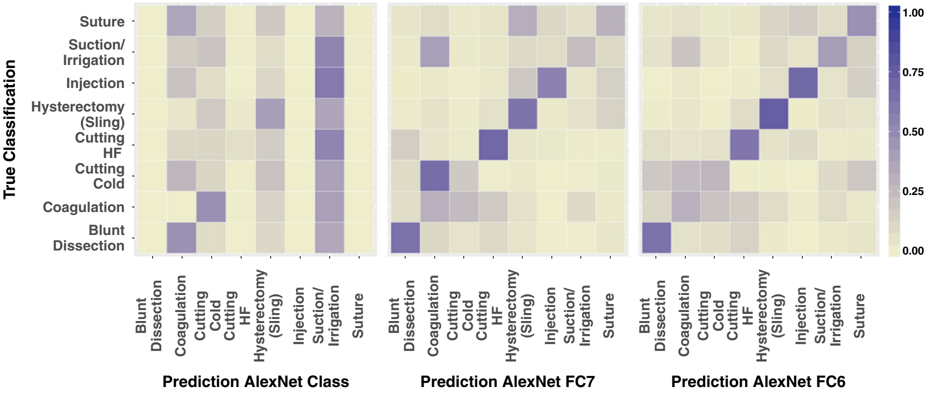


Fig. 7 Confusion matrices for SVM action classification using AlexNet features

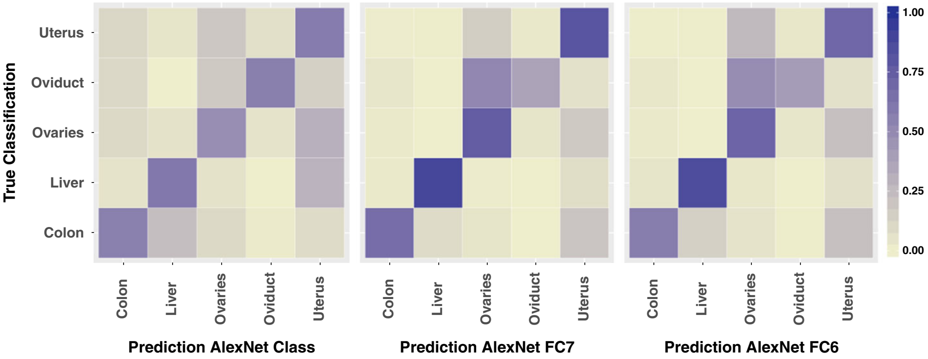


Fig. 8 Confusion matrices for SVM anatomy classification using AlexNet features

We visualize the performance for our surgical action and anatomical structure classes of the individual approaches using confusion matrices depicted in Fig. 5 for the CNN action models, and Fig. 6 for CNN anatomy models. SVM confusion matrices are given in Fig. 7 for action classification, and Fig. 8 for anatomy classification. Columns denote the predicted class while rows indicate the true class. Cell shades illustrate prediction percentage relative to the number of examples for a class.

CNN and SVM action models perform poorest in the classes *Coagulation*, *Cutting Cold*, and *Suction & Irrigation*. We think this originates in the fact that the single-frame CNN models have limited means to model the way the instruments are used. For CNN and SVM anatomy models, there is a bias to confuse the classes *Ovaries*, *Uterus* and *Oviduct*. We think this originates in the fact that these organs are spatially very near and when these organs are on the images, it is likely that parts of those other classes are visible as well.

6 Conclusion

In this paper, we investigate CNN.based single-frame classification models for video shots in gynecological surgery. Together with medical experts, we provide a first taxonomy for important anatomical structures and surgical actions of interest for the domain of laparoscopy videos in gynecology. For this domain, we build a dataset of 9 h of video data manually extracted from 111 different medical interventions. In particular, we train two different CNN architectures AlexNet and GoogLeNet from scratch for both, surgical action and anatomical structure classification. Furthermore, we investigate an SVM approach using off-the-shelf neural network features from AlexNet: *class*, *fc7*, and *fc6*. The best results from the SVM approach using features extracted from AlexNet using off-the-shelf weights outperform the full AlexNet CNN trained from scratch in both, anatomical structure as well as action classification which might originate in the choice to label the database scene-wise and not on a per-frame basis. Moreover, GoogLeNet, the best-performing approach on general images, also is the best performing approach in this domain. These results imply that advances in general image classification domains can lead to advances in difficult expert domains, such as our use case of gynecological surgery video classification.

Despite the fact that this domain is pretty narrow, there is plenty of future work to do. We think a per-pixel classification approach for anatomical structures could yield more accurate results for structures which are spatially near each other. More examples for future work include the evaluation of more sophisticated approaches for video classification, such as frame fusion models or LSTM-based models. Also, the question of whether we can surpass human performance by adding more network depth remains open. However, we think that classification of surgical actions provides the most benefit for surgeons and therefore focus on the following point. We assume that the capabilities of the used single-frame CNN models AlexNet and GoogLeNet are not fully utilized. Hence, we aim at an improvement of surgical action classification by using early fusion of raw image data with multiple (domain-specific) modalities of which at least one represents a temporal dimension, such as motion vectors.

Acknowledgements Open access funding provided by University of Klagenfurt. This work was supported by Universität Klagenfurt and Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF 20214 u. 3520/ 26336/38165.

References

1. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N (2016) Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans Med Imaging 35(5):1313–1321. doi:10.1109/TMI.2016.2528120
2. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S (2016) Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 35(5):1207–1216. doi:10.1109/TMI.2016.2535865
3. Beecks C, Schoeffmann K, Lux M, Uysal MS, Seidl T (2015) Endoscopic video retrieval: a signature-based approach for linking endoscopic images with video segments. In: Del Bimbo A, Chen SC, Wang H, Yu H, Zimmermann R (eds). IEEE, Los Alamitos, pp 1–6
4. Dixit M, Chen S, Gao D, Rasiwasia N, Vasconcelos N (2015) Scene classification with semantic fisher vectors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2974–2983
5. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia, MM '14. ACM, New York, pp 675–678. doi:10.1145/2647868.2654889
6. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: The IEEE conference on computer vision and pattern recognition (CVPR)
7. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Bartlett P, Pereira F, Burges C, Bottou L, Weinberger K (eds) Advances in neural information processing systems 25, pp 1106–1114
8. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M (2014) Medical image classification with convolutional neural network. In: 2014 13th international conference on control automation robotics & vision (ICARCV). IEEE, pp 844–848
9. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. arXiv:1702.05747
10. Loukas C, Nikiteas N, Schizas D, Georgiou E (2016) Shot boundary detection in endoscopic surgery videos using a variational bayesian framework. Int J Comput Assist Radiol Surg 11(11):1937–1949. doi:10.1007/s11548-016-1431-2
11. Münzer B, Schoeffmann K, Böszörmenyi L (2013) Relevance segmentation of laparoscopic videos. In: 2013 IEEE international symposium on multimedia, pp 84–91. doi:10.1109/ISM.2013.22
12. Münzer B, Schoeffmann K, Böszörmenyi L (2017) Content-based processing and analysis of endoscopic images and videos: a survey. Multimed Tools Appl 1–40. doi:10.1007/s11042-016-4219-z
13. Münzer B, Schoeffmann K, Böszörmenyi L, Smulders JF, Jakimowicz JJ (2014) Investigation of the impact of compression on the perceptional quality of laparoscopic videos. In: Krol M (ed) 27th international symposium on computer-based medical systems (CBMS'14). IEEE, New York City, pp 153–158
14. OpenCV (2015) Open Source Computer Vision Library. https://github.com/itseez/opencv
15. Park SY, Sargent D (2016) Colonoscopic polyp detection using convolutional neural networks. In: SPIE medical imaging. International Society for Optics and Photonics, pp 978528–978528
16. Petscharnig S, Schöeffmann K (2017) Deep learning of shot classification in gynecologic surgery videos. In: Amsaleg L, Guðmundsson G, Gurrin C, Jónsson B, Satoh S (eds) MultiMedia Modeling. MMM 2017. Lecture Notes in Computer Science, vol 10132. Springer, Cham
17. Primus MJ, Schoeffmann K, Böszörmenyi L (2015) Instrument classification in laparoscopic videos. In: 2015 13th international workshop on content-based multimedia indexing (CBMI), pp 1–6. doi:10.1109/CBMI.2015.7153616
18. Primus MJ, Schoeffmann K, Böszörmenyi L (2016) Temporal segmentation of laparoscopic videos into surgical phases. In: 2016 14th international workshop on content-based multimedia indexing (CBMI), pp 1–6. doi:10.1109/CBMI.2016.7500249
19. Qiu Y, Wang Y, Yan S, Tan M, Cheng S, Liu H, Zheng B (2016) An initial investigation on developing a new method to predict short-term breast cancer risk based on deep learning technology. SPIE Medical Imaging. International Society for Optics and Photonics, pp 978521–978521

20. Quellec G, Lamard M, Cochener B, Cazuguel G (2014) Real-time segmentation and recognition of surgical tasks in cataract surgery videos. IEEE Trans Med Imaging 33(12):2352–2360. doi:10.1109/TMI.2014.2340473
21. Ribeiro E, Uhl A, Wimmer G, Häfner M (2016) Transfer learning for colonic polyp classification using off-the-shelf cnn features. In: International workshop on computer-assisted and robotic endoscopy. Springer, pp 1–13
22. Samala RK, Chan HP, Hadjiiski LM, Cha K, Helvie MA (2016) Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis. In: SPIE medical imaging. International Society for Optics and Photonics, pp 97850Y–97850Y
23. Schoeffmann K, Del Fabro M, Szkaliczki T, Böszörmenyi L, Keckstein J (2014) Keyframe extraction in endoscopic video. Multimedia Tools and Applications, pp 1–20. doi:10.1007/s11042-014-2224-7
24. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298. doi:10.1109/TMI.2016.2528162
25. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: The IEEE conference on computer vision and pattern recognition (CVPR)
26. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: full training or fine tuning IEEE Trans Med Imaging 35(5):1299–1312. doi:10.1109/TMI.2016.2535302
27. 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 Trans Med Imaging 36(1):86–97. doi:10.1109/TMI.2016.2593957
28. Xing F, Xie Y, Yang L (2016) An automatic learning-based framework for robust nucleus segmentation. IEEE Trans Med Imaging 35(2):550–566. doi:10.1109/TMI.2015.2481436
29. Yan Z, Zhan Y, Peng Z, Liao S, Shinagawa Y, Zhang S, Metaxas DN, Zhou XS (2016) Multi-instance deep learning: discover discriminative local anatomies for bodypart recognition. IEEE Trans Med Imaging 35(5):1332–1343. doi:10.1109/TMI.2016.2524985
30. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: The IEEE conference on computer vision and pattern recognition (CVPR)

Stefan Petscharnig started his studies of computer science in 2010 at AAU Klagenfurt. He received his bachelor's degree in 2014 and his master's degree in 2015 from AAU Klagenfurt. In his master's thesis he investigated the impact of asynchronism on the Quality of Experience in social TV like scenarios using methods from Crowdsourcing, Human Computation, and Gamification. He was occupied as student research assisstant and after graduation as project assistant for the AdvUHD-DASH project in 2015. Since 2016, he works in the KISMET research project with a focus on the analysis of endoscopic video data using machine learning.

Dr. Klaus Schöffmann is an Associate Professor in the distributed multimedia systems research group at the Institute of Information Technology (ITEC) at Klagenfurt University, Austria. He received his Ph.D. in 2009 and his Habilitation (venia docendi) in 2015, both in Computer Science and from Klagenfurt University. His research focuses on human-computer interaction with multimedia data (e.g., video browsing), multimedia content analysis, and multimedia systems (particularly in the mobile and medical domain). He has co-authored more than 80 publications on various topics in multimedia and he has co-organized international conferences, special sessions and workshops (e.g., MMM 2012, CBMI 2013, VisHMC 2014, MMC 2014, MMC 2015, and MMC 2016). He is co-founder of the Video Browser Showdown (VBS), an editorial board member of the Springer International Journal on Multimedia Tools and Applications (MTAP), Springer International Journal on Multimedia Systems, and a steering committee member of the International Conference on MultiMedia Modelling (MMM). Additionally, he is member of the IEEE and the ACM and a regular reviewer for international conferences and journals in the field of multimedia. Prof. Schoeffmann teaches various courses in computer science, including mobile app development (interactive multimedia applications), video retrieval, media technology, distributed systems, distributed multimedia systems, and operating systems.