Robustness Against Out of Distribution Video Frames in Online Surgical Workflow Recognition with Temporal Convolutional Networks

Amirhossein Bayat[1] AMIRHOSSEIN.BAYAT@CARESYNTAX.COM
Kadir Kirtac, Salih Karagoz, Julien Schwerin, Michael Stenzel, Marco Smit
Florian Aspart[1] FLORIAN.ASPART@CARESYNTAX.COM
[1] *Caresyntax, Berlin, Germany*

Abstract

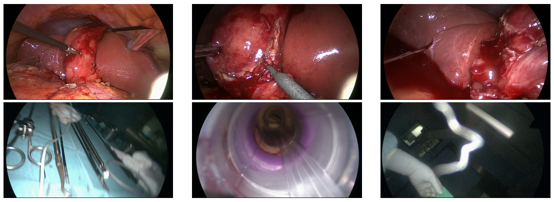
The automatic recognition of surgical phase based on laparoscopic videos is a pre-requisite to diverse AI application on surgeries. Online surgical phase recognition is commonly achieved using two-stages models combining (i) a spatial feature extraction at the frame level with a (ii) temporal model. Yet, this online surgical phase recognition is a challenging task in real-world scenarios. For example, the camera might be temporally extracted of the body during surgeries (e.g., to be cleaned). The Out-of-body (OOB) phases have out-of-distribution spatial features and have unpredictable occurrence which affect the temporal model performance. We propose a simple, yet effective, mechanism to robustify our temporal model against OOB phases. Our solution leverages the two-stages structure of surgical phase model predictions. We train and test our model on a large scale real-world dataset of laparoscopic cholecystectomy videos and show the effectiveness of our approach.

Keywords: Temporal Convolutional Networks, Surgical Workflow, OOD Detection.

1. Introduction

Over the last decades, laparoscopic surgeries have gained an increasing popularity. The camera inserted in the abdomen, opened the door to video-based, automatic surgical phase recognition (Czempiel et al., 2020). These deep learning studies are mainly based on the proposed approaches for action recognition models in computer vision (Farha and Gall, 2019; Li et al., 2020). These approaches rely on two-stage models: (i) a spatial feature extraction model at the frame level, followed (ii) by a temporal model. Due to the different modality of surgical videos, new challenges have to be resolved. Surgical videos tend to be long, with a limited the field of view and variable lightning condition. Additionally the different objects have similar texture. (see the first row in Figure 1 for some examples). Another challenge for intra-operative, online surgical phase recognition is the presence of out-of-body (OOB) phases. These phases have random occurrence, e.g., when the camera is extracted to be cleaned, and variable length, such as when facing a longer task with no instrument movements. Combined with their random temporal occurrence, the out-of-distribution spatial features of OOB frames (compared to inside body) can be problematic for online intra-operative phase recognition models. In this work, we propose to leverage the two-step characteristic of phase detection models to handle the OOB frames. Specifically, we trained the spatial feature extraction model to concurrently detect OOB frames. The temporal model is then automatically bypassed for these given frames. We show that our approach outperforms the state of the art in online phase prediction by a good margin.

Figure 1: Examples of (top) inside body and (bottom) out of body (OOB) frames.



2. Our method

OOB frames represent a small portion of the entire dataset. Yet, the variability of OOB environment is much higher than what is observed in the rest of the dataset. In fact, while inside body frames display similar textures and color range, OOB images contain totally different scenes and objects Figure 1. This is reflected by the channel-wise pixel intensity distribution of the OOB frames, which is wider and with different mean compared to inside body frames in Figure 2. While the intensity distribution follow an almost normal distribution in inside body frames (blue), the OOB (red) distribution is considerably different. More importantly, OOB sequences can appear randomly during the surgeries, which complicates the tasks for the model to learn the natural transition between surgical phases. We argue that detecting OOB frames and excluding them from the sequential data could improve the model performance and robustness.

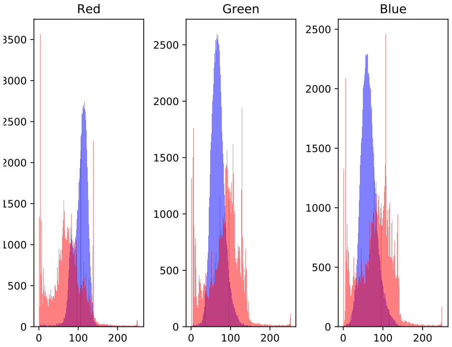


Figure 2: Inside/OOB intensity distribution

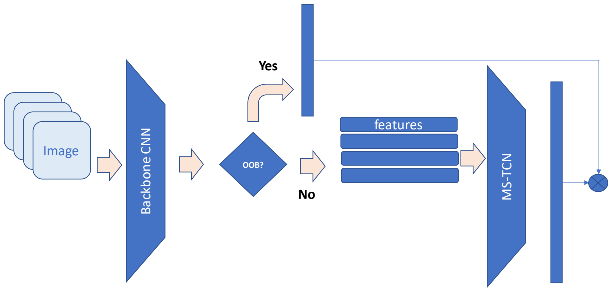


Figure 3: Proposed workflow

Based on the ideas mentioned above, we design a workflow for processing the surgical videos, depicted in Figure 3. At the core of our pipeline lies a multistage temporal convolutional (MS-TCN) network for surgical phase recognition. MS-TCN consists of a serie of dilated 1D convolutions, which enables the model to have a large temporal receptive field with less parameters (Farha and Gall, 2019). This network takes the spatial features extracted for each frame by our backbone network, a Resnet-50 model. The extracted feature vectors are added to a feature array in chronological order and then fed to the temporal model. Concurrently to the spatial feature extraction, the OOB frames are detected by the backbone network and excluded from MS-TCN's input. We train the Resnet-50 model using two heads: one for predicting the surgery phase and one for detecting the OOB frames. For the OOB frames detection head, we define an out-of-distribution(OOD) problem by using a one class classifier. Inside body frames are positive examples and OOB frames negative ones. In our temporal network we define four stages with Dilated Residual Layers. To increase the receptive field, we employ Dual Dilated Residual (DDL) layers in the first stage (Li et al., 2020). Since we are employing our model in an online application, we use causal convolutions. To train this network we employed weighted cross entropy loss for each stage.

Dataset. We evaluated our approach on an in-house dataset with 307 videos. All possible video recordings, including additional procedures, emergency surgery or teaching cases, were included in our dataset.We shuffled the dataset and selected 70, 10, 20 percentage of the video data for training, validation and testing respectively.

2.1. Results and Discussion

Method	Accuracy	F1-Score	Precision	Recall
MS-TCN	0.86±0.10	0.67±0.20	0.71±0.20	0.71±0.21
MS-TCN++	0.87±0.09	0.67±0.21	0.71±0.21	0.69±0.23
OOB-bypass	0.88±0.8	0.69±0.22	0.73±0.19	0.73±0.23
OOB-bypass+MH-Resnet	0.89±0.09	0.70±0.21	0.74±0.19	0.71±0.22

Table 1: Comparing our approach with other methods for phase detection.

In this section, we evaluate the proposed method and compare it to other models in Table 1. All models are trained on our dataset with presence of OOB frames. As can be seen, our method for detecting and bypassing OOB frames operates better than the other methods MS-TCN(Farha and Gall, 2019) and MS-TCN++(Li et al., 2020). First we test our approach for bypassing the OOB frames by detecting them using Resnet-50 model. The Resnet model is trained to predict the surgical phase given a single frame. According to the predicted phase OOB frames could be detected. After that, we train the multihead Resnet model (MH-Resnet) as explained in the methods section. By using features extracted with MH-Resnet and bypassing the OOB frames we get the best performance. Our proposed approach is easy to implement and could be used with existing temporal models, including CNNs, LSTMs and Transformers.

References

Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pages 343–352. Springer, 2020.

Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019.

Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2020.