EndoRCN: Recurrent Convolutional Networks for Recognition of Surgical Workflow in Cholecystectomy Procedure Video

Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, and Pheng-Ann Heng

Dept. of Computer Science and Engineering, The Chinese University of Hong Kong

1 Introduction

Automated analysis of surgical workflow has been playing an increasingly important role in modern operating room, since the surgical workflow would directly affect the patient safety and the high-stress working conditions of clinicians [5]. Recognition of the surgical phase has been gradually becoming an essential basis for monitoring the surgery process, scheduling surgeons, predicting upcoming events, alerting and suggesting modifications, etc [4]. In addition, automatic segmentation of the surgical video can also help to facilitate the surgeon skill evaluation and improve the efficiency of documenting surgical reports.

Previous studies have relied on signals of surgical tools to carry out the cholecystectomy procedure phase recognition in laparoscopic videos [1]. However, the information about tool usage relies heavily on manual annotation or additional built-in sensors [3]. Recognising surgical phase solely based on visual information is a promising direction, yet a quite challenging task. The main challenges lie in the high scene variability and blur within the surgery videos due to the blood occlusion and camera motion. Twinanda et al. [5] have utilized a convolutional neural network (CNN) consisting of 9 layers to learn features from visual information of the frames, and achieved state-of-the-art performance.

In this work, we propose to exploit a very deep residual convolutional network (50 layers) to effectively extract highly discriminative features purely from visual information of the video. Furthermore, based on the visual representations from the convolutional network, we propose to encode the temporal information with a long-short term memory (LSTM) network, which is crucial for video analysis tasks. Overall, we name our proposed framework of recurrent convolutional networks for surgical flow recognition as *EndoRCN*.

2 Method

2.1 Frame visual feature extraction with residual CNN

The residual networks have demonstrated outstanding accuracy and compelling optimization behaviors on large-scale image recognition tasks, such as ImageNet [2]. Considering the challenges in the surgical phase recognition task, we propose to exploit the deep residual network for robust visual feature extraction within each single frame. In terms of architecture, the network is composed of a set of residual blocks, each of which consists of a few stacked layers, including convolution layers, batch normalization layers and relu layers. The residual learning is performed via shortcut connections and element-wise addition, which enable direct gradients propagation from one block to another.

We first downsampled the original 25fps video into 5fps and resized them into the resolution of 250×250. The images were further augmented with 227×227 cropping and mirroring before input to the CNN. During training, our residual convolutional network was initialized from the ResNet-50 model which consisted of 50 layers and was pre-trained on the large-scale ImageNet dataset [2]. We modified the last fully connected layer into 8 neurons to match the classes of surgical phases in our task.

2.2 Temporal information modeling with LSTM

We observe that the temporal information within the video is crucial for recognizing the surgical phases, since the intra-class variations are huge and different phases might have some similar frame scenes. In this regard, we propose to rely on the LSTM units for encoding the temporal information of the video. The key point is that the LSTM needed inputting previous frames, so that the hidden neurons in the network can store and update temporal information. Thus, the insight to include the recurrent neural network lies in that the model is able to not only consider the current frame, but also what it has perceived from previous frames in earlier time. Since we processed the video in the online mode, our model did not look at frames afterwards.

2.3 EndoRCN: Recurrent Convolutional Networks

Our framework integrates the deep residual network and an LSTM unit for the surgical phase recognition task. In practice, we first downsampled the original 25fps video into 5fps and change the resolution of the static image to 250×250. After several augmentation operations, we input them to the convolutional networks which have been pre-trained to get the feature information of each frame. Then, given the current frame f_t, we combined its former several frames f_j ($j < t$) to form a short-period video clip. Specifically, we included the current frame and its former two frames to construct each video clip, considering the rapidness of key action during the surgery. In experiments, we trained two models, one sampling preceding frames at stride of 3 (i.e., f_{t-6}, f_{t-3}, f_t), another sampling former frames at stride of 2 (i.e., f_{t-4}, f_{t-2}, f_t). Each of the three frame in a video clip was first input to the CNN and we obtained a 2048-dimensional visual feature vector for each single frame. Next, the three feature vectors corresponding to the three frames were sequencially input to the LSTM for modeling temporal information, and output the phase prediction of frame f_t. The number of hidden states in the LSTM unit was 512, and we averaged the prediction probabilities of the two models to yield the EndoRCN results.

2.4 Postprocessing

We finally conducted simple yet effective online postprocessing to further enhance the temporal consistency with preceding surgical workflow, by leveraging the natural characteristics of the specific task. Our strategy was inspired by the important observation of the relatively fixed order within the workflow. For example, it was impossible that a frame of phase 'TrocarPlacement' following a frame of phase 'CalotTriangleDissection'; and similarly, the phase 'Gallbladder-Packaging' was unable to happen behind the phase 'GallbladderRetraction'.

In practice, we first investigated which phase has begun at the current time, by counting the number of previous frames which were continuously classified as a phase P by the EndoRCN. We assumed that the current frame belonged to the phase P, when the counting number was greater than a threshold. Then, if the current frame was classified as neither the phase P nor the next phase $P{+}1$, we repurposed the current frame as phase P. In addition, we further observed that sometimes a frame would be misclassified into the next phase, and this problem can not be handled with the aforementioned strategy. Fortunately, we noticed that the EndoRCN prediction probability in this kind of situation was not that high, which meant that the model was not very confident to classify the frame into the next phase. In this regard, we further set a prediction probability threshold, lower than which the phase recognition output would be amended as the phase P. Overall, we postprocessed the video frame by frame from the beginning to the end, by only using the phase predictions of previous frames and the current frame prediction probability. We did not look at subsequent frames to ensure our recognition method performed in the online mode.

The last important thing is that, the high-quality results from EndoRCN is crucial for the postprocessing strategy, because it heavily relied on the high consistency and accurate prediction probability from the network.

3 Learning Process

We first pre-trained the deep residual convolutional network using 1300k static images extracted from the video frames. The training batch size was set to 10. The learning rate started from 0.0005 and was divided by 10 every 20k iterations, and the weight decay was set to 0.005. After that, the CNN and the LSTM unit were trained simultaneously in an end-to-end manner towards the phase label of the input frames. Different from previous step, the CNN was set smaller learning rate and our main objective here was to learn the LSTM unit. We implemented the deep learning framework based on the *Caffe* library. It took around 1.5 days to train the whole *EndoRCN* framework with a GPU of Nvidia Titan Z.

4 Results

We evaluated our framework on the surgical workflow dataset which M2CAI challenge, one of MICCAI 2016 challenges provided. Our work is one of the top three performing methods and yields a jaccard score of 78.2.

References

1. Blum, T., Feußner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 400–407. Springer (2010)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
3. Lin, H.C., Shafran, I., Murphy, T.E., Okamura, A.M., Yuh, D.D., Hager, G.D.: Automatic detection and segmentation of robot-assisted surgical motions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 802–810. Springer (2005)
4. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Medical image analysis 16(3), 632–641 (2012)
5. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. arXiv preprint arXiv:1602.03012 (2016)