Endo3D: Online Workflow Analysis for Endoscopic Surgeries Based on 3D CNN and LSTM

Weixiang Chen[1,2,3], Jianjiang Feng[1,2,3](✉), Jiwen Lu[1,2,3], and Jie Zhou[1,2,3]

[1] Department of Automation, Tsinghua University, Beijing, China
jfeng@tsinghua.edu.cn
[2] State Key Lab of Intelligent Technologies and Systems,
Tsinghua University, Beijing, China
[3] Beijing National Research Center for Information
Science and Technology, Beijing, China

Abstract. Surgical workflow analysis is an important topic of computer-assisted intervention and phase recognition is one of its important tasks. Features extracted from video frames by 2D convolutional networks were proved feasible for online phase analysis in former publications. In this paper, we propose to extract fine-level temporal features from video clips using 3D convolutional networks (CNN) and use Long Short-Term Memory (LSTM) networks to capture coarse-level information. By combining fine-level and coarse-level information, our proposed method outperforms state-of-the-art online methods without using specific knowledge of surgeries and almost reaches the state-of-the-art offline performance.

1 Introduction

Computer-assisted surgery system (CAS) is an important topic of computer-assisted intervention, which assists surgeons by giving some advice or guidances in surgeries. To achieve this aim, Surgeries Workflow Analysis (SWA) is an important task. Endoscopic surgery workflow analysis progresses rapidly these years because this kind of surgeries are all performed under an endoscopic cameras so that the videos are always available. In addition, endoscopic surgeries need CAS more than other surgeries because of the limited field of view in endoscopic camera. With such limited field of view, it is very difficult for surgeons to recognize the detailed positions of the camera, the targets, and some special vessels or nerves.

Existing publications on SWA have described various types of features which can be roughly divided into image-based features and signal-based features. Signal-based features are extracted from signals like tool usage [14], some manually defined surgical activities [11], and kinematic data [13]. Although signal-based features yield good performance, it requires some additional devices (e.g. RFID tags for tool signals and daVinci system for kinematic data), which is inconvenient for many online situations. Since surgery videos are always available,

D. Stoyanov et al. (Eds.): OR 2.0/CARE/CLIP/ISIC 2018, LNCS 11041, pp. 97–107, 2018.
https://doi.org/10.1007/978-3-030-01201-4_12

image-based features can be more universal. At first, image-based features were mainly extracted by manually designed rules. [2] used pixels value and its gradients; [4] designed descriptors combining color, shape and texture information. However, manually selected image features are suboptimal.

A better solution to this problem is selecting features by convolutional neural networks (CNN) instead of manually. With appropriate setup, CNN can learn highly distinctive features from training data. EndoNet [15] used AlexNet to extract features and fed them to hybrid Hidden Markov Models (HMM) for phase recognition. On the dataset of the EndoVis 2015 Workflow Challenge, EndoNet performed the best. A recent method SV-RCNet [9], which combines ResNet [7] and Long Short-Term Memory (LSTM) [6], is now the state-of-the-art on Cholec80 dataset[1] [15].

EndoNet used AlexNet as the basic network and extracted features from a single frame. This limited the expressiveness of features because they contain no temporal information. SV-RCNet used LSTM to mix shot features into clip features, but since all convolutions were still in single shot, it ignores edges in time domain. EndoNet used HMM for global optimization which performs well in its offline version. However, online analysis is important in many applications, such as giving doctors some advice during surgery or in emergency situations. SV-RCNet's LSTM method can work online, but clips of 2s are too short to cover coarse-level temporal features. Without prior knowledge inference (PKI) which is specific to certain surgeries [9], its accuracy is 85.3%, only slightly higher than EndoNet's.

We proposes an online SWA method Endo3D which is based on C3D networks [8] and LSTM. It extracted 3-D CNN features from a clip of video rather than a single frame, which encodes fine-level temporal information. Besides, we proposed a three-layer LSTM with sequences long enough to encodes coarse-level temporal information into our prediction. Our proposed method outperforms SV-RCNet (whose accuracy without PKI is 85.3%) in online recognition with 91.2% online accuracy on Cholec80. In addition, it can also predict tool usage with 86% Mean Average Precision (mAP). The main contributions of our method are:

1. Extract spatial-temporal features from surgery videos with an extended C3D network.
2. Extract coarse-level information by LSTM which plays important role in phase recognition.
3. Combine fine-level and coarse-level temporal information in an online mode.
4. Achieve state-of-the-art online phase recognition accuracy without using specific knowledge.
5. Achieve high accuracy in tool presence detection.

2 Methodology

2.1 Endo3D Network Architecture

Our model is trained in two steps (as Fig. 1 shows). The first step is fine tuning process on a network derived from C3D. We use the fine tuned network to extract

[1] http://camma.u-strasbg.fr/datasets.

features and predict tool presences. In the second step, the features are arranged into sequence which is then be fed to 3-layer LSTM to predict workflow phases. For every time step in sequence, our model gives a prediction for phases. The two parts of Endo3D separately introduce fine-level (for about 4.6 s) and coarse-level (for all the past frames) temporal information into recognition, together with spatial information.

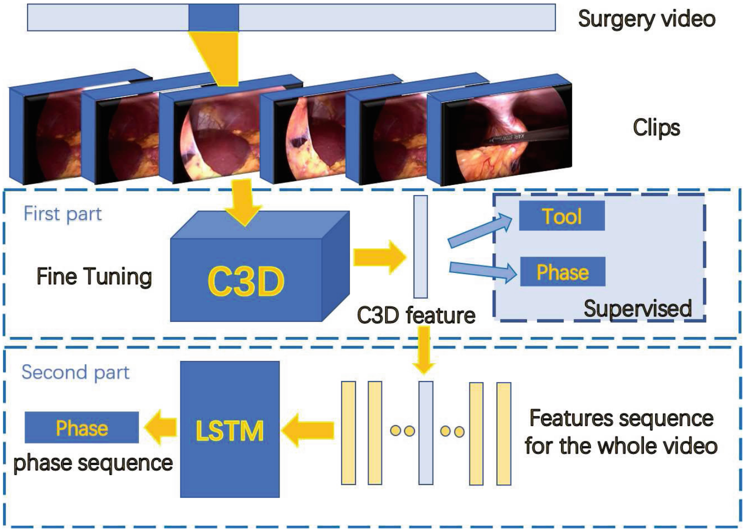


Fig. 1. Diagram of the proposed Endo3D method.

2.2 First Part: Video Feature Extraction

The first part of our model is shown in Fig. 2. The C3D's fc7 layer is supposed to compute tool presence. There is a concatenation layer fc8 after fc7 which concatenates tool layer and fc7 for phase prediction. After training, the phase and tool layer are left away and we use fc8 as a $l_v = 4103$ dimensional feature. In other words, phase layers are only used as supervisions in training. The input of our network are $16 \times 112 \times 112 \times 3$ RGB video clips and the output feature vectors is denoted as V_f. We downsample videos from 25 fps to 2.5 fps, and arrange contiguous 16 frames as a clip in length of 4.6 s and with a sampling interval of 1 s. As a result, fine-level temporal texture is introduced when doing three-dimensional convolution in this step.

This part is trained using Adam [12]. Our tool layer's output is activated by sigmoid function, because tool presence detection is a multi-labeled task. We write it as V_t whose length is the number of tools denoted as n_t. V_p denotes the phase layer's output which is activated by softmax function. For a batch of size N, loss function can be defined as:

$$L = c_1 \times L_t + c_2 \times L_p + c_3 \times L_{regu} + c_4 \times L_w \tag{1}$$

$$L_t = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{n_t}[T_t^{(i)}\log(V_t^{(i)}) + (1 - T_t^{(i)})\log(1 - V_t^{(i)})] \tag{2}$$

$$L_p = -\frac{1}{N}\sum_{i=1}^{N}\sum_{p=1}^{n_p} T_p^{(i)}\log(V_p^{(i)}) \tag{3}$$

where T_p and T_t is groundtruth of phase and tool respectively; c_is are weighting coefficients; L_w is weight decay loss and L_{regu} is regularization loss, which are set to prevent overfitting; (i) means the i-th sample in the batch.

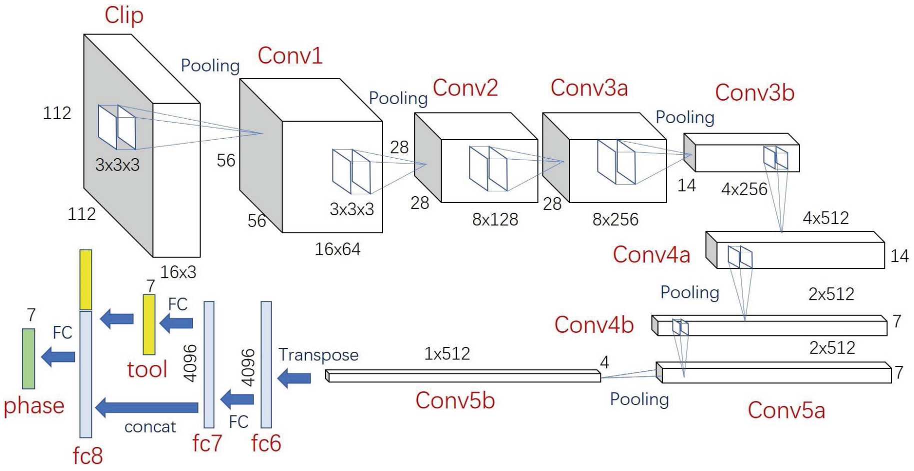


Fig. 2. The first part: C3D network of the proposed method.

2.3 Second Part: Coarse-Level Temporal Information

Since C3D captures only fine-level temporal information we introduce LSTM to deal with long term temporal information, which is shown in Fig. 3. fc8 vectors are arranged into sequences and fed to LSTM. For every time step in the sequence, the output will only be influenced by all the past inputs, so our method is online. $V_{f,t} \in \mathbb{R}^{\mathrm{l_v} \times 1}$ denotes the value of fc8 layer of t-th timestep (the outputs of former networks are strided with 1 s, so the timestep of LSTM is 1 s). The sequence is denoted as $S^{(T)} = [V_{f,1}, V_{f,2}, ..., V_{f,T}]$, where T denotes the sequence at T-th second. Because LSTM networks care nothing about the length of sequences, we use feature sequence of all past clips as input and get output of the same number of clips. Only the output of last timestep is used as the newest coming prediction in testing procedure.

In order to simplify training, we expand all sequences to the same length of n_s with 0 and set their labels with background class which is different from

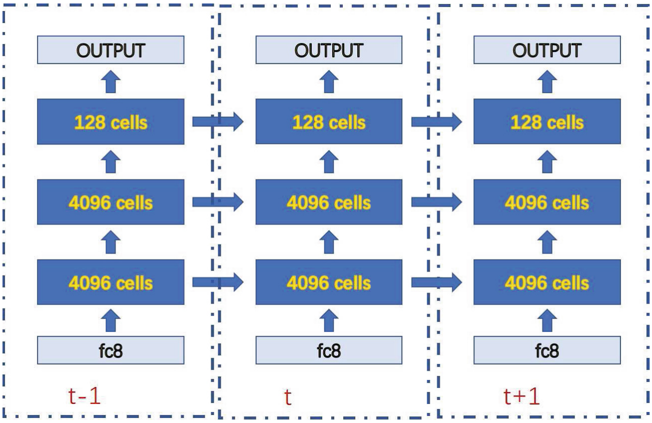


Fig. 3. The second part: LSTM network of the proposed method

real phases. We also introduce a mask to mark background frames. Expanded sequence is denoted as $\widehat{S^{(T)}} = [S^{(T)}, O_{l_v, n_s - T}]$ and all the sequences are now in the same shape of $l_v \times n_s$. Output sequences are $n_p \times n_s$ dimensional binary vectors, which is denoted as $P_{t,p}$ corresponding to phase p and timestep t. Mask is also a vector denoted as M_t. $M_t = 0$ if timestep t is background, otherwise $M_t = 1$.

The sequence learning loss can be denoted as:

$$L_s = -\frac{1}{N}\sum_{i=1}^{N}\frac{\sum\limits_{t=1}^{n_s}\sum\limits_{p=1}^{n_p} M_t^{(i)} T_{t,p}^{(i)} \log(P_{t,p}^{(i)})}{\sum\limits_{t=1}^{n_s} M_t^{(i)}}. \tag{4}$$

It is a cross-entropy loss with mask to filter out background. When computing accuracy, background results are not taken into account. Output of LSTM can be directly used after softmax function as the confidence value without other classifiers.

We trained the whole network from scratch. We choose 3-layer LSTM because the number of its parameters is suitable for the difficulty of the problem and the size of training set. If new SWA tasks are defined, the complexity of this part can be changed.

3 Experiment

3.1 Dataset

Experiments are done on Cholec80 dataset [15], which contains 80 videos of cholecystectomy surgeries performed by 13 surgeons. All the videos are captured at 25 fps and are sampled to 2.5 fps. The whole set is labeled with tool presences and phases. Video frames are annotated with 7 phases (see Fig. 4) in 2.5 fps. Phases are notated as P0 to P6 following the order above. For most videos, transformations between phases follow some disciplines. Unlike Jin *et al.* [9], we do not use this prior knowledges. Tools are annotated in 1 fps which also have 7 kinds[2].

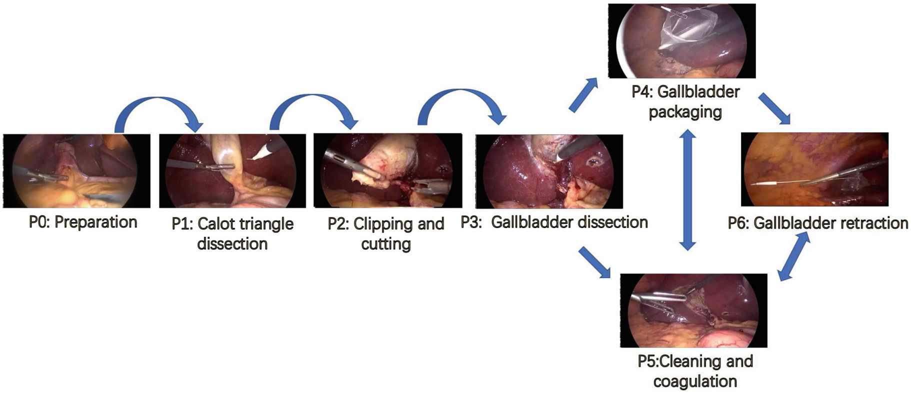

P0: Preparation
P1: Calot triangle dissection
P2: Clipping and cutting
P3: Gallbladder dissection
P4: Gallbladder packaging
P5:Cleaning and coagulation
P6: Gallbladder retraction

Fig. 4. Surgery workflow of cholecystectomy for dataset Cholec80.

For the first part, we use first 40 videos to train Endo3D feature extraction network and the other 40 videos for validation and test. For the second part, 40 validation videos are divided to do 4-fold cross-validation, which is the same as the division of EndoNet. In training set, there are over 200K frames with their annotations. We arrange them into 16 frames length clips with 1 s stride and finally get about 86K clips. The labels of clips are defined by label of the last frame in the clip, because we want to use only past frames to extract features. In the second step, videos are transformed into feature sequences and fit into n_s, the length of the longest one in our dataset. Only complete sequences are used in training.

3.2 Training Parameters

Our C3D network is pretrained on sport1M dataset [10]. The fc7 and fc8 layers are trained at the learning rate of 10^{-3} and initialized randomly. The layers defined in original C3D networks are initialized using pretrained parameters and trained at 10^{-4}. Training for this part is setup on 2 NVIDIA GeForce 1080Ti cards. The batch size is 24 per card. Our process is carried out using Tensorflow [1] and training process takes 16 h for all 10K iterations. Feature extraction takes approximately 156 ms per clip on one card.

For the last 40 videos of Cholec80 dataset, $n_s = 5983$. The output of LSTM network is 8D feature vector, because $n_p = 7$ for Cholec80. We trained our

[2] Seven tools: Bipolar, Clipper, Grasper, Hook, Irrigator, Scissors, Specimen bag.

model for 80 epochs with batch size 2 and initial learning rate 0.01. LSTM process is carried out on Keras [3] and executed on a 1080Ti card. Training process takes approximately 90 s for an epoch and it takes about 2 s to predict a sequence, which is related to n_p. Because Keras implementation can not predict dynamically with timesteps, for every new coming timestep, it computes from the first timestep to the last one and only updates the new coming timestep's prediction. Changing some Keras' backend code or implementing the method with C language can accelerate this process.

3.3 Results

Phase Recognition. Phase recognition is measured by precision, recall, and accuracy which are defined in [13]. The results are shown in Tables 1 and 2. Results of EndoNet and SC-RCNet are cited from the reference paper [9,15]. Notations of all baselines are defined as follow:

- **EndoNet SVM**: EndoNet [15] without its HMM, which are the recognition results feeding fc8 of EndoNet into SVM.
- **EndoNet ON**: the online phase recognition results of EndoNet [15]
- **EndoNet OFF**: the offline phase recognition results of EndoNet [15].
- **SV-RCNet+PKI**: the phase recognition result of SV-RCNet with prior knowledge inference process [9].
- **SV-RCNet**: the phase recognition result of SV-RCNet without prior knowledge inference process [9].
- **C3D**: the results of our phase layer's output fine-tuned with only phase supervision.
- **Endo3D**: results of our phase layer's output fine-tuned with proposed tool and phase supervisions.
- **Endo3D SVM**: results of our fc8 after a SVM classifier.
- **Endo3D LSTM**: results of our proposed Endo3D process.

Table 1. Phase recognition results (%).

Method	Precision	Recall	Acc.
EndoNet no-HMM	70.1	66.7	75.3
EndoNet ON	75.1	80.0	81.9
EndoNet OFF	85.7	89.1	92.2
SV-RCNet+PKI	90.6	86.2	92.4
SV-RCNet	80.7	83.5	85.3
C3D	63.5	59.9	69.9
Endo3D	66.4	67.0	74.7
Endo3D SVM	72.8	68.4	78.7
Endo3D LSTM	81.3	87.7	91.2

Table 2. Compare for every phase on precision and recalls (%).

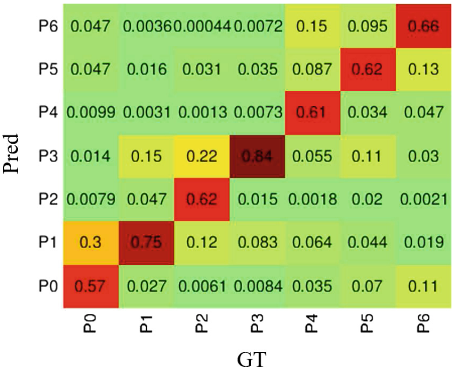
Phase	Method (Precision/Recall)	
ID	EndoNet ON	Endo3D LSTM
P0	**90.0**/85.5	82.8/**99.8**
P1	96.4/81.1	**96.9/97.8**
P2	**69.8/71.2**	69.5/71.0
P3	82.8/86.5	**97.3/88.8**
P4	55.5/75.57	**92.3/91.7**
P5	**63.9**/68.7	58.2/**81.6**
P6	57.5/**88.9**	**72.1**/82.5

Endo3D with LSTM outperforms other online methods and is almost comparable to offline version of EndoNet. Only for some short phases like P0, P2 and P6, the proposed method does not perform as well as EndoNet. The result without LSTM is comparable to EndoNet no-HMM method (which uses SVM), and our Endo3D SVM method outperformed it. SV-RCNet without prior knowledge is not as good as our method and our method can almost reach its result with prior knowledge. C3D features perform a little worse than the proposed method, which proves that using tool information as supervision in training and as features in predicting phases has positive influences.

LSTM in our method and HMM in EndoNet can both improve results a lot. According to our result, the contribution of LSTM is greater than HMM. Theoretically, HMMs are based on transition matrix, emission matrix, whose representation ability may be lower than LSTM. LSTM use forget gates to manage memories from far before, which improves performances in long sequences learning. Besides, LSTM can be easily extended to multi-layers.

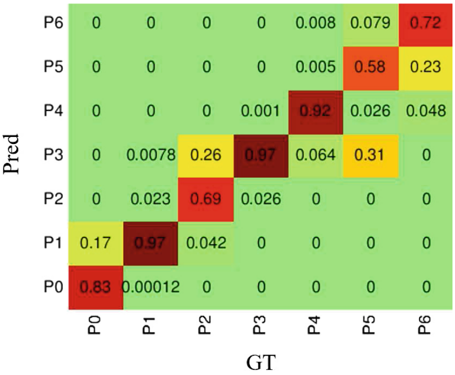
Prior Knowledge Inference (PKI) helps SV-RCNet a lot in accuracy, but we suppose that such knowledge should be better learnt by network from videos. As an automatic method, prior knowledge for specific dataset might not always be available. Data-driven methods can be extended to new surgery datasets without manually defined knowledge, which we suppose is a desirable property.

Figures 5 and 6 show the confusion matrix of 7 phases and the background for C3D features without and with LSTM, respectively. Predictions spread on less phases after LSTM, which shows LSTM does help filter out impossible transformations. P5 is the only phase getting worse after LSTM and it is predicted as P3 for many cases. As Fig. 4 shows, P5 is next to P3, P4 or P6 which is the most complex phase from the perspective of coarse-level phase transformations. Irrigators are mainly used in P5 which is detected with high accuracy, so from the perspective of tool evidences, P5 is not that difficult and our prediction before LSTM is a little higher.



	P0	P1	P2	P3	P4	P5	P6
P6	0.047	0.0036	0.00044	0.0072	0.15	0.095	0.66
P5	0.047	0.016	0.031	0.035	0.087	0.62	0.13
P4	0.0099	0.0031	0.0013	0.0073	0.61	0.034	0.047
P3	0.014	0.15	0.22	0.84	0.055	0.11	0.03
P2	0.0079	0.047	0.62	0.015	0.0018	0.02	0.0021
P1	0.3	0.75	0.12	0.083	0.064	0.044	0.019
P0	0.57	0.027	0.0061	0.0084	0.035	0.07	0.11

Fig. 5. Confusion matrix before LSTM.



	P0	P1	P2	P3	P4	P5	P6
P6	0	0	0	0	0.008	0.079	0.72
P5	0	0	0	0	0.005	0.58	0.23
P4	0	0	0	0.001	0.92	0.026	0.048
P3	0	0.0078	0.26	0.97	0.064	0.31	0
P2	0	0.023	0.69	0.026	0	0	0
P1	0.17	0.97	0.042	0	0	0	0
P0	0.83	0.00012	0	0	0	0	0

Fig. 6. Confusion matrix after LSTM.

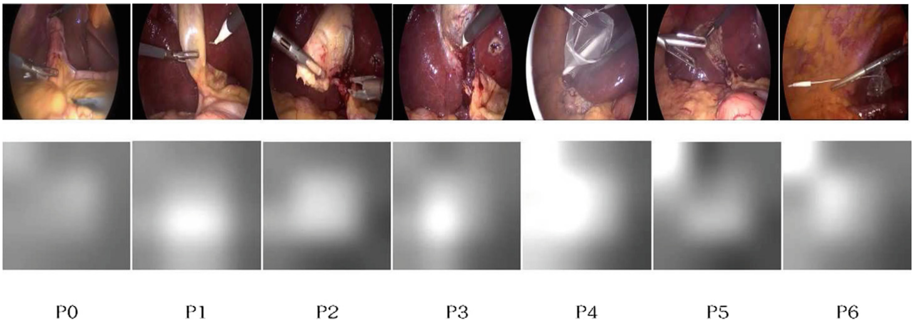


Fig. 7. Average feature maps from Conv5b layer.

Figure 7 shows average feature maps extracted by the proposed C3D networks. The maps come from the last pooling layer of network and are average between channels. We arrange feature maps according to their groundtruth phases as Fig. 7 shows. Eventhough it is hard to describe the detailed meanings of deep features, we can find out that feature maps have different reaction regions for different phases.

Table 3. Tool presence detection result (%).

Tool	DPM	EndoNet	Endo3D
Bipolar	60.6	**86.9**	69.72
Clipper	68.4	80.1	**95.12**
Grasper	82.3	**84.8**	71.32
Hook	93.4	**95.6**	87.81
Irrigator	40.5	74.4	**96.43**
Scissors	23.4	58.6	**87.33**
Specimen bag	40.0	86.8	**94.97**
MEAN	58.4	81.0	**86.1**

Tool Presence Detection. The tool presence performance is measured by mAP. Results about EndoNet are reported in [15]. Deformable Part Model (DPM) [5], one of the most popular object detection method, is used as a baseline for tool presence detection.

The results are shown in Table 3. The mAP for Bipolar, Grasper and Hook of proposed method is lower than EndoNet, but for the other 4 tools its mAP is higher. For Irrigator, Scissors and Clipper, the mAPs are higher for more than 15% points. As a result, average mAP for all tools of our proposed method is about 5 points higher than EndoNet. In fact, Grasper and Hook might occur in almost all phases, because surgeons need them to move or grasp tissues. So these two tools are less important as phase features. The proposed method is more sensitive to tools like scissors and irrigator, whose occurrences are key information for phase, because we train tool detection together with phase recognition.

4 Conclusions

In this paper, we focus on online phase recognition of endoscopic surgery videos and propose a method to learn 3-D CNN features from video clips called Endo3D. With the help of C3D and LSTM network, we combine fine-level and coarse-level temporal texture together and use temporal-spatial information to recognize phases. In addition, Endo3D uses tool and phase groundtruth to do multi-target training. The proposed method outperformed the previous state-of-the-art on public domain dataset without using specific knowledge.

Reducing the time consumption is the first thing to do in the future. As an online method, the current processing time limits the output rate. Keras consumes most of time because this implementation doesn't support dynamical input and output of LSTM nodes. Engineering improvements like a C version test script will help a lot because average time to compute per node of LSTM is less than 40 ms.

Acknowledgement. This work is supported by the National Natural Science Foundation of China under Grant 61622207.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). Software available from http://tensorflow.org/
2. Blum, T., Feußner, H., Navab, N.: Modeling and segmentation of surgical workflow from laparoscopic video. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6363, pp. 400–407. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15711-0_50
3. Chollet, F., et al.: Keras (2015). https://github.com/keras-team/keras
4. Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., Jannin, P.: Automatic data-driven real-time segmentation and recognition of surgical workflow. IJCARS **11**(6), 1–9 (2016)
5. Felzenszwalb, P.F., Girshick, R.B., Mcallester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE T-PAMI **32**(9), 1627 (2010)
6. Graves, A.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
8. Ji, S., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE T-PAMI **35**(1), 221–231 (2012)
9. Jin, Y., et al.: SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. IEEE T-MI **37**(5), 1114–1126 (2018)

10. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F.F.: Large-scale video classification with convolutional neural networks. In: CVPR, pp. 1725–1732 (2014)
11. Katić, D., et al.: Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 158–167. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07521-1_17
12. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. Comput. Sci. (2014)
13. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. Med. Image Anal. **16**(3), 632–641 (2012)
14. Stauder, R., et al.: Random forests for phase detection in surgical workflow analysis. In: Stoyanov, D., Collins, D.L., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI 2014. LNCS, vol. 8498, pp. 148–157. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07521-1_16
15. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., Mathelin, M.D., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. IEEE T-MI **36**(1), 86–97 (2016)