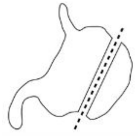
SWNet: Surgical Workflow Recognition with Deep Convolutional Network

Bokai Zhang, Amer Ghanem, Alexander Simes, Henry Choi, Andrew Yoo, Andrew Min C-SATS, Inc., Johnson & Johnson

1. Purpose

1.1 Project Background:
Video-based automatic surgical workflow recognition is one of the key technologies to build computer-assisted interventional systems for modern operating rooms.(1) Early studies propose to use CNN and RNN [1] or CNN and Multi-Stage Temporal Convolutional Network (MS-TCN) [2] to solve the problem. (2) We propose to use deep 3DCNN, MS-TCN, and a post-process algorithm to solve the problem.



1.2 Medical Background:
Sleeve Gastrectomy is used to assist patients with losing excess weight.
Eight surgical phases: (1) Exploration (2) Ligation of short gastric vessels (3) Gastric transection (4) Bougie (5) Suturing of omentum to stomach (6) Liver retraction (7) Hiatal hernia repair and (8) Gastric band removal. The parts of the video that did not get annotated were named as (9) Not a surgical phase.

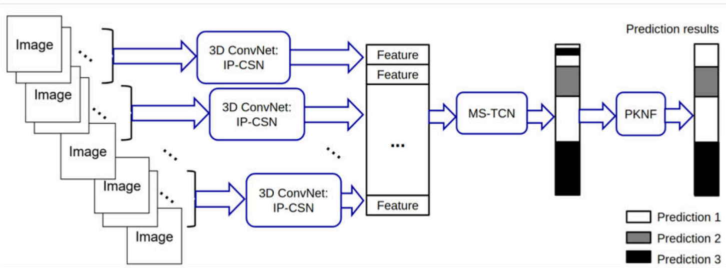
2. Dataset and Method

2.1 Dataset:
317 videos to train.
82 videos to validate.
62 videos to test.

Table 3: Training, validation and test datasets (minutes of video)

Phase Name	Training Data	Validation Data	Testing Data
Not a surgical phase	5729.91	1460.91	1202.01
Ligation of short gastric vessels phase	4247.63	1082.03	828.13
Gastric transection phase	3988.37	953.85	690.50
Bougie phase	305.08	64.35	50.62
Suturing of omentum to stomach phase	2562.70	807.70	397.62
Exploration phase	181.83	38.33	27.22
Liver retraction phase	65.48	25.97	6.88
Hiatal hernia repair phase	448.95	72.38	102.63
Gastric band removal phase	52.63	42.32	31.03

2.2 Method Summary for SWNet:
- Divide the full surgery video into short video segments. Use Interaction-Preserved Channel-Separated Convolutional Network (IP-CSN [3]) to extract features for each video segment.
- Combine the segment-level features and use MS-TCN [4] to achieve initial surgical phase segmentation for the full video.
- We apply the Prior Knowledge Noise Filtering (PKNF) algorithm to the initial surgical phase segmentation results to get the final prediction results for the full video.



We fine-tune IP-CSN (Pretrained on IG65M [5] and Kinetics-400 [6]) on our dataset.
We develop PKNF in consideration of three aspects: phase order, phase incidence, and phase time.

3. Results

3.1 Offline recognition results:
- Different methods are compared: ResNetLSTM [1], TeCNO(ResNet-MSTCN) [2], EfficientNet-MSTCN with/without PKNF, IPCSN-LSTM with/without PKNF, and IPCSN-MSTCN with/without PKNF.
- Table 1 shows that: (1) IP-CSN is a better feature extraction backbone. (2) MS-TCN is a better video action segmentation network. (3) Adopting PKNF can reduce noise and improve prediction results. (4) SWNet outperforms all other approaches.

Table 1: Overall accuracy and Jaccard score for offline surgical workflow recognition

Method	Accuracy	Weighted Jaccard Score
ResNetLSTM	0.8235	0.7141
TeCNO	0.8659	0.7668
EfficientNet-MSTCN	0.8818	0.7928
EfficientNet-MSTCN-PKNF	0.8861	0.7995
IPCSN-LSTM	0.8548	0.7505
IPCSN-LSTM-PKNF	0.8713	0.7744
IPCSN-MSTCN	0.8921	0.8070
IPCSN-MSTCN-PKNF (SWNet)	0.9037	0.8256

- As shown in Figure, we visualize the prediction results for 4 test videos. It is clear that our SWNet can locate the surgical phase more accurately and identify phase transactions better compares to other methods.

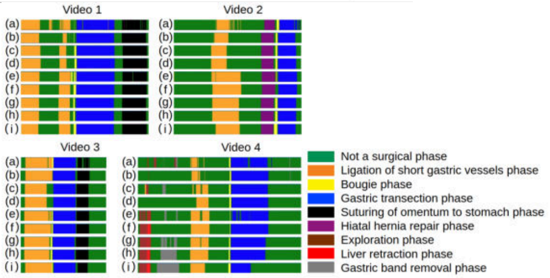


Figure 2: Color-coded ribbon illustration for offline recognition results: (a) ResNetLSTM prediction results (b) TeCNO prediction results (c) EfficientNet-MSTCN model output (d) EfficientNet-MSTCN-PKNF prediction results (e) IPCSN-LSTM model output (f) IPCSN-LSTM-PKNF prediction results (g) IPCSN-MSTCN model output (h) SWNet prediction results (i) Ground Truth

3.1 Online recognition results:
- As shown in the Table, our IPCSN-MSTCN trained with smooth loss significantly outperforms other methods from segmental evaluation metric aspects.
- As shown in the Figure, our method has fewer over-segmentation errors and out-of-order predictions comparing to other approaches.

Table 2: Overall accuracy, segmental edit distance and segmental F1 for online surgical workflow recognition

Method	Accuracy	Jaccard	Edit	F1@10	F1@25	F1@50
ResNetLSTM	0.8130	0.6997	22.2775	23.2044	20.6931	15.7710
TeCNO	0.8451	0.7331	42.5531	46.7005	43.8578	35.7360
IPCSN-MSTCN(L_{cls})	0.8425	0.7326	49.5681	49.6224	44.8759	33.6570
IPCSN-MSTCN($L_{cls} + \lambda L_{T-MSE}$)	0.8466	0.7367	56.5213	56.1170	52.9255	41.4894

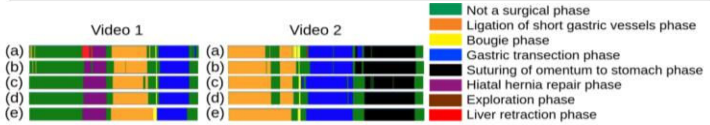


Figure 3: Color-coded ribbon illustration for online recognition results: (a) ResNetLSTM prediction results (b) TeCNO prediction results (c) Predictions from IPCSN-MSTCN trained with L_{cls} (d) Predictions from IPCSN-MSTCN trained with $L_{cls} + \lambda L_{T-MSE}$ (e) Ground Truth

4. Conclusion

In this paper, we designed SWNet for surgical workflow recognition with IP-CSN, MS-TCN, and PKNF. For both online and offline surgical workflow recognition, our SWNet outperforms several other approaches and can achieve state-of-the-art results.

Reference

[1] Jin, Yueming, et al. "SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network." *IEEE transactions on medical imaging* 37.5 (2017): 1114-1126.
[2] Czempiel, Tobias, et al. "TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2020.
[3] Tran, Du, et al. "Video classification with channel-separated convolutional networks." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
[4] Farha, Yazan Abu, and Jurgen Gall. "Ms-tcn: Multi-stage temporal convolutional network for action segmentation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
[5] Ghadiyaram, Deepti, Du Tran, and Dhruv Mahajan. "Large-scale weakly-supervised pre-training for video action recognition." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
[6] Kay, Will, et al. "The kinetics human action video dataset." *arXiv preprint arXiv:1705.06950* (2017).