2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Spatio-Temporal Features Based Surgical Phase Classification Using CNNs

Chakka Sai Pradeep[1] and Neelam Sinha[1]

Abstract— **In this paper, we propose a novel encoder-decoder based surgical phase classification technique leveraging on the spatio-temporal features extracted from the videos of laparoscopic cholecystectomy surgery. We use combined margin loss function to train on the computationally efficient PeleeNet architecture to extract features that exhibit: (1) Intra-phase similarity, (2) Inter-phase dissimilarity. Using these features, we propose to encapsulate sequential feature embeddings, 64 at a time and classify the surgical phase based on customized efficient residual factorized CNN architecture (ST-ERFNet). We obtained surgical phase classification accuracy of 86.07% on the publicly available Cholec80 dataset which consists of 7 surgical phases. The number of parameters required for the computation is approximately reduced by 84% and yet achieves comparable performance as the state of the art.**

Clinical relevance— **Autonomous surgical phase classification sets the platform for automatically analyzing the entire surgical work flow. Additionally, could streamline the process of assessment of a surgery in terms of efficiency, early detection of errors or deviation from usual practice. This would potentially result in increased patient care.**

I. Introduction

Laparoscopic cholecystectomy is a minimally invasive gall bladder removal surgery where surgical instruments are inserted into the abdomen through small incisions with the help of a laparoscope. With increasing importance of minimally invasive surgeries there is a spurt in data availability. As the availability of information has increased, surgical video analysis has become important to improve the overall patient care. Surgical phase classification is an important aspect in optimizing the entire surgical work flow.

A systematic review of surgical phase classification using ML techniques was provided in [7], which dealt with surgeries that could have varying number of phases. However, first large scale cholecystectomy dataset was released along with the works of EndoNet [8]. In their works, they had used AlexNet based feature extractor, SVM classifier along with HMM to exploit the temporal constraint on the surgical work flow. SV-RCNet [10] is a retrospective study, where the authors trained an end to end ResNet-LSTM network which requires prior knowledge about surgery duration for developing the inference system. Same authors in their work of MTRCNet [11] approached surgical phase classification as a multi-tasking problem. Here, surgical tool features were extracted and were fed to a LSTM model for surgical phase recognition. Temporal convolutions [13] were used in TeCNO [9] for the first time for surgical phase classification. Authors had used causal, dilated multi-stage temporal convolution networks in their work which had achieved the then current state of the art results. We report our results in comparison with this benchmark since same dataset is utilized.

II. Proposed Method

The purpose of this study is the surgical phase classification, towards achieving that contributions of this paper are: (i) To use encoded features of preceding and current frames to classify the phase of cholecystectomy surgery in real-time i.e. to use a causal CNN approach. (ii) To use a computationally efficient network architecture without any significant reduction in accuracy. We achieve a reduction of 84% in number of parameters with comparable performance as the state of the art. (iii) To use combined margin loss for the first time for surgical feature embedding.

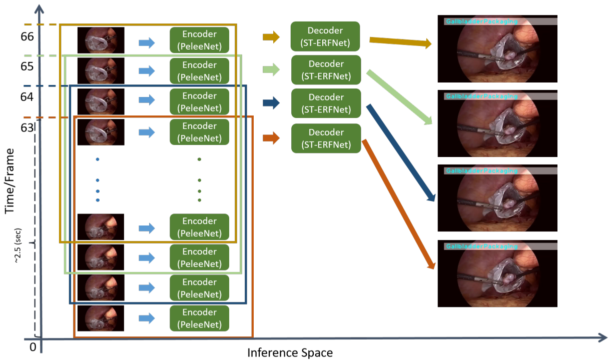

Fig. 1. System Overview

In this study, we propose the usage of Encoder-Decoder CNN system architecture in causal framework, represented in Fig. 1. Encoder follows PeleeNet [1] architecture and Decoder is a customized form of ERFNet [2] which we refer to as ST-ERFNet from here on. Table I gives an overview of proposed decoder architecture. Complete proposed architecture details of training and deploy networks along with other ablation studies have been provided in the following repository: https://github.com/csai-arc/SPR-peleenet-custom_erfnet.

Fig. 2 describes the utilized PeleeNet architecture which takes an RGB image (which is a snapshot of time freezed moment of a surgery) of size $3 \times 240 \times 427$ and produces encoded feature tensor of dimension 1×128 from fully connected layer. 64 such sequential feature embeddings are

[1] Chakka Sai Pradeep and Neelam Sinha are with International Institute of Information Technology, Bangalore, India. saipradeep.chakka@iiitb.ac.in, neelam.sinha@iiitb.ac.in

stacked together and is given as input to decoder network. Decoder provides a probabilistic score of predicted surgical phase for each of the 64 frame embeddings tensor of dimension $1 \times 64 \times 128$.

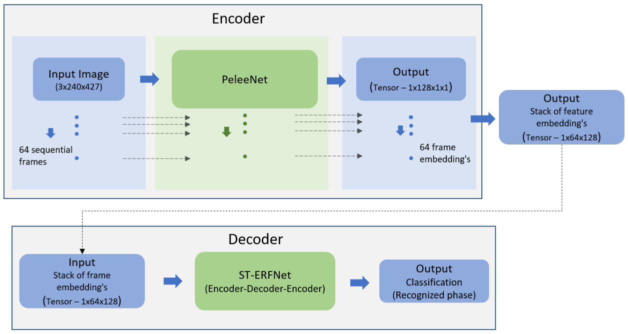

Fig. 2. System Architecture

In this work, we chose 64 frames which approximates to 2.5 seconds of video feed in real-time. Embeddings of these past 63 frames along with the current frame enables our system to make a robust history based classification on the surgical phase. It is empirically seen that human action with a tool cannot change within a finer time frame.

In Table I, Column **Stage** represents the stage in which a particular block is present. For each layer/block, **L** represents the number, **In-Res** denotes the input resolution, **n** denotes the number of repetitions, **s** represents the overall stride, **Out-Res** denotes the output resolution.

A. *Encoder Architecture*

The already existing PeleeNet architecture, with conventional convolution operation is used as the encoder. This enables usage of available hardware optimized libraries for real-time image feature embeddings needed as input for our decoder. As described in [1] stem block is the first layer connected to data layer before the first dense layer used in PeleeNet network architecture. This block enhances the feature expression ability. Dense block used is a two way dense layer to obtain distinct scales of receptive features. There are three distinct connections in a dense block: (a) identity/skip connection, (b) convolution with 3×3 kernel size and (c) two stacked convolution with kernel size 3×3. These three connections enable the network to learn visual cues in larger field of view. Transition Block used is a convolution layer having the same number of output features as that of input. Dense Blocks and transition blocks are used in sequence to produce an encoded feature tensor of dimension 1×128 from fully connected layer. 64 such sequential feature embeddings are stacked together and is given as input to decoder network.

B. *Decoder Architecture*

A novelty of the proposed study lies in customizing the ERFNet architecture to incorporate Spatio-Temporal information, hence called ST-ERFNet. ERFNet [2] was originally introduced to achieve real-time semantic segmentation on natural images. In this study, we design an Encoder-Decoder-Encoder architecture which efficiently mixes the spatio-temporal feature embeddings. These embeddings are obtained by stacking current frame embedding along with the preceding 63 frame embeddings to classify the surgical phase. Improved prediction of the surgical phase is made possible attributing to this efficient combination of Spatio-Temporal features. We exploit the three available dimensions within a CNN architecture. i.e feature space, temporal space and spatial space. Input layer dimension of the proposed ST-ERFNet architecture is $1 \times 64 \times 128$, 1 being feature space, 64 being temporal space and 128 spatial space. In the first encoder stage, learning is enforced on the feature space by setting a dimension of $128 \times 8 \times 16$. Decoder stage enforces learning on the spatio-temporal domain, second encoder stage enforces learning again on the feature domain. In this way, we blend spatio-temporal features and classify the surgical phase.

TABLE I
ST-ERFNET ARCHITECTURE OVERVIEW

Stage	L	In-Res	Type	n	s	Out-Res
Encoder	1	1x64x128	Downsampler Block	1	2	16x32x64
	2	16x32x64	Downsampler Block	1	2	64x16x32
	3-5	64x16x32	NBD Block	3	1	64x16x32
	6	64x16x32	Downsampler Block	1	2	128x8x16
	7-11	128x8x16	NBD Block	5	1	128x8x16
Decoder	12	128x8x16	Upsampler Block	1	2	64x16x32
	13-15	64x16x32	NBD Block	3	1	64x16x32
	16	64x16x32	Upsampler Block	1	2	16x32x64
	17-19	16x32x64	NBD Block	3	1	16x32x64
Encoder	20	16x32x64	Downsampler Block	1	2	64x16x32
	21-23	64x16x32	NBD Block	3	1	64x16x32
	24	64x16x32	Downsampler Block	1	2	128x8x16
	25-27	128x8x16	NBD Block	3	1	128x8x16
Convolution	28	128x8x16	Convolution Layer	1	1	128x8x16
Classification	29	128x8x16	Global pooling Layer	1		128x1x1
	30	128x1x1	Fully Connected Layer	1		7
	31	7	Softmax	1		7

Table I describes the various blocks used in our ST-ERFNet architecture. Downsampler block performs downsampling by concatenating parallel outputs of convolution with kernel size 3×3 and max pooling with kernel size 2×2 with stride 2. This block reduces the spatial resolution because of its stride 2. However, this allows deeper layers to learn more contextually and also reduces the overall computation. Non-bottleneck-1D block (NBD), first introduced in [2], factorizes residual layers with no bottle necks. Two serial convolution operations of kernel sizes 3×1 and 1×3 are used instead of a single convolution with kernel size 3×3 along with identity/skip connection. Hence, number of training parameters are reduced. Upsampler blocks use simple deconvolution layers with stride 2. Layers 1 to 11 comprise the first encoder segment, layers 12-19 comprise the decoder segment and layers 20-27 comprise the second encoder segment whose output is used for softmax surgical phase classification.

III. Dataset and Training Methodology

Proposed methodology has been illustrated on Cholec80 dataset introduced in [8]. This dataset contains 80 cholecystectomy surgeries performed by 13 surgeons at the University hospital of Strasbourg. The annotated surgical videos are available at http://camma.u-strasbg.fr/datasets. In this study, the first 64 have been used for training and the remaining for testing (80%-20% split). There are seven classes of surgical phases, they are: preparation, calot triangle dissection, clipping cutting, gallbladder dissection, gallbladder packaging, cleaning coagulation and gallbladder retraction.

Encoder Training: In this work, it is proposed to utilize **combined margin loss** defined in the equation below, for the first time in the application of surgical phase classification. The encoder is trained over combined margin loss, which was originally proposed for the face recognition task in [6]. The advantage of this loss function is that it ensures intra-phase similarities and inter-phase dissimilarities on the extracted feature embeddings over different surgical phases.

$$L = -\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{s(cos(m_1\theta_{y_i}+m_2)-m_3)}}{e^{s(cos(m_1\theta_{y_i}+m_2)-m_3)}+\sum_{j=1,j\neq y_i}^{n} e^{scos\theta_j}}$$

Where m_1, m_2 and m_3 denote the margin penalties: SphereFace [5], ArcFace [4] and CosFace [6] are hyper-parameters that are set to 1.0, 0.3 and 0.2 respectively. θ_{y_i} is the angle between j^{th} weight and i^{th} feature. The learned embeddings are thus distributed on a hypersphere with a radius s set to 64. N is the batch size. Angular margin penalties are included to simultaneously enhance the intra-phase similarities and inter-phase dissimilarities. Encoder has been trained using Adam solver with momentum 0.9 and momentum2 of 0.999. Multistep learning rate policy with base learning rate of 0.0001. Training over 100,000 iterations with a batch size of 35 was done using this setup.

ST-ERFNet Decoder Training: The input to the ST-ERFNet decoder is 64 feature embeddings obtained from the encoder, while the output is the classification label of the surgical phase. The inferred feature embeddings from the encoder are stacked for 64 sequential images and dumped into HDF5 packets as described in the Fig. 2 for all the available sequences in the dataset. Hence, temporal information is inducted along with the already existing spatial information for Decoder training. 64 sequential frames which approximately encapsulate 2.5 seconds of data in real-time are used to ensure appropriate mixing of information across the three dimensions of feature, space and time. The proposed ST-ERFNet decoder network architecture in section II-B is trained over softmax loss. SGD solver with a momentum of 0.9, Triangular learning rate policy [3] with base learning rate of 0.0001, max learning rate of 0.002 and step size value of 2000 has been used for training with a batch size of 180.

IV. Results and Discussions

TABLE II
Evaluation Results Comparison

Method	Accuracy	Precision	Recall
Binary tool [8]	47.5 ± 2.6	54.5±32.3	60.2 ± 23.8
Handcrafted [8]	32.6 ± 6.4	31.7 ± 20.2	38.4 ± 19.2
Handcrafted+CCA [8]	38.2 ± 5.1	39.4 ± 31.0	41.5 ± 21.6
AlexNet [8]	67.2 ± 5.3	60.3 ± 21.2	65.9 ± 16.0
PhaseNet [8]	78.8 ± 4.7	71.3 ± 51.6	76.6 ± 16.6
EndoNet [8]	81.7 ± 4.2	73.7 ± 16.1	79.6 ± 7.9
PhaseLSTM [9]	79.68±0.07	72.85±0.10	73.45±0.12
EndoLSTM [9]	80.85±0.17	76.81±2.62	72.07±0.64
MTRCNet [11]	82.76±0.01	76.08±0.01	78.02±0.13
ResNetLSTM [9]	86.58±1.01	80.53±1.59	79.94±1.79
TeCNO [9]	88.56±0.27	81.64±0.41	85.24±1.06
Our Method	86.07±0.04	77.48±0.05	72.19±0.07

Proposed methodology achieves an accuracy of 86.07% with just over 4.36M parameters. Table II provides a performance comparison between our proposed system against the state of the art systems in terms of accuracy, precision and recall. Encoder is designed with 2.187M parameters and decoder with 2.174M parameters. This method outperforms most LSTM based methods which can be computationally demanding. Best performace was reported in [9] which inherently utilizes temporal convolution network proposed in [13]. [9] utilizes ResNet50 architecture for feature extraction which itself contains 23.7M parameters followed by 2 stage TCN [13] architecture which approximately contains 2M parameters. An approximate 26M parameters were needed to achieve an accuracy of 88.56%. Our proposed method utilizes just 16% of this total number of parameters in comparison with the state of the art methodology with just 2.5% trade off in terms of accuracy.

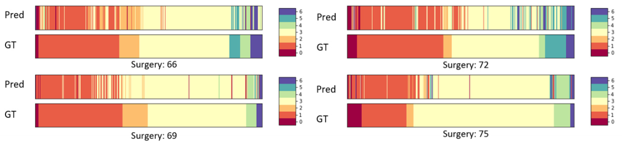

Fig. 3. Surgical phase classification results vs ground truth time progression comparison on Cholec80 dataset over surgical videos 66, 69, 72 and 75 are shown here. Color coding is done in the following way: 0-Preparation, 1-CalotTriangleDissection, 2-ClippingCutting, 3-GallbladderDissection, 4-GallbladderPackaging, 5-CleaningCoagulation and 6-GallbladderRetraction.

Fig. 3 illustrates color-coded ribbon plot comparison between predicted surgical phases and ground truth. Unique color code is assigned to each phase for visual analysis.

It was observed that in reality during the phase of "calot triangle dissection" there may arise a need of partial "gall bladder dissection". However, such intricate details were not available as part of ground truth labels. Due to this, these two classes are more prone to confusion. Similarly, the hard separation between "preparation" phase and "calot triangle dissection" phase is subjective since different doctors annotate differently. Fig. 4 depicting the confusion matrix of the predicted phases validates our observations. Few of the test scenarios are illustrated in the Fig. 5.

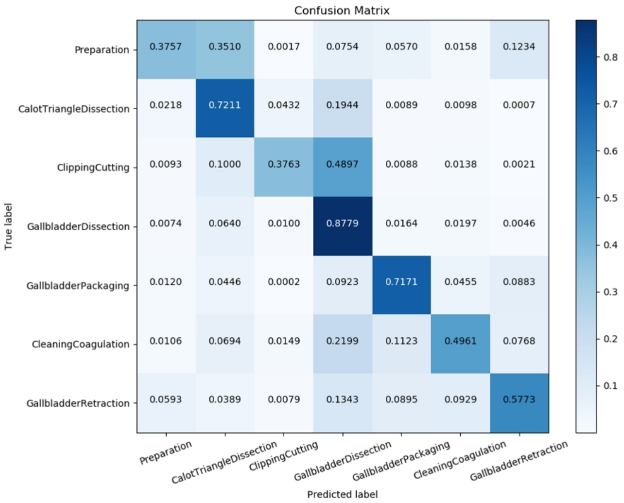

Fig. 4. Confusion Matrix of the proposed Surgical Phase classification system

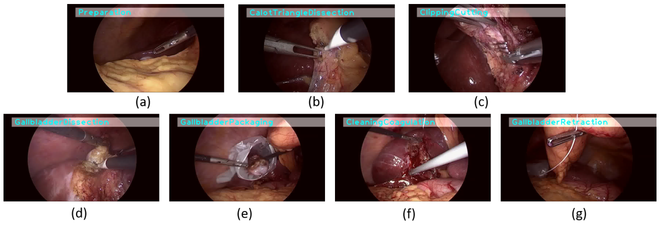

Fig. 5. Performance of our proposed system on few of the laparoscopic cholecystectomy surgical scenarios is shown here: (a) preparation, (b) calot triangle dissection, (c) clipping cutting, (d) gallbladder dissection, (e) gallbladder packaging, (f) cleaning coagulation and (g) gallbladder retraction.

Cholec80 dataset provides surgical tool usage labels at 1fps for each surgery. Further experiments can be conducted for classification of surgical tools along with surgical phase in a multi-tasking manner which could lead to improved performance, based on reported works on real world images [14]. The misclassification between "clipping cutting" and "gall bladder dissection" is 48.97%, perhaps due to similar tools at similar locations in these two phases. Also, "gall bladder dissection" phase is maximally available, leading to to an overall false positive rate of 0.21.

V. Conclusion

We propose a real-time causal Encoder-Decoder CNN architecture for surgical phase classification on laparoscopic cholecystectomy surgical videos. To incorporate spatio-temporal information we propose a novel Spatio-Temporal ERFNet (ST-ERFNet), which serves as decoder, while existing PeleeNet architecture is used as encoder. Encoder was trained to extract distinct features over combined margin loss. Encoded features of the preceding 63 frames and feature embeddings of the current frame were used to classify the phase, here illustrated on publicly available laparoscopic cholecystectomy surgery dataset. The proposed approach achieves accuracy of 86.07% with just over 4.36M parameters which is 84% reduction in terms of number of parameters with respect to state of the art.

References

[1] Wang, Robert J. and Li, Xiang and Ling and Charles X., (Jan. 2019). "Pelee: A Real-Time Object Detection System on Mobile Devices," [Online]. Available: https://arxiv.org/abs/1804.06882

[2] Eduardo Romera, José M. Álvarez, Luis M. Bergasa, and Roberto Arroyo, "ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation," *IEEE Transactions on Intelligent Transportation Systems,* vol. 19, no: 1, pp. 263 - 272, Jan 2018.

[3] Leslie N. Smith, (Apr. 2019). "Cyclical Learning Rates for Training Neural Networks," [Online]. Available: https://arxiv.org/abs/1506.01186

[4] Jiankang Deng, Jia Guo, Niannan Xue and Stefanos Zafeiriou, (Feb. 2019). "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," [Online]. Available: https://arxiv.org/abs/1801.07698

[5] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj and Le Song, (jan. 2018). "Sphereface: Deep hypersphere embedding for face recognition," [Online]. Available: https://arxiv.org/abs/1704.08063

[6] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, Wei Liu, (Apr. 2018). "Cosface: Large margin cosine loss for deep face recognition," [Online]. Available: https://arxiv.org/abs/1801.09414

[7] Garrow CR, Kowalewski KF, Li L, Wagner M, Schmidt MW, Engelhardt S, Hashimoto DA, Kenngott HG, Bodenstedt S, Speidel S, Müller-Stich BP, Nickel F. "Machine Learning for Surgical Phase Recognition: A Systematic Review," *Annals of surgery,* vol. 273, no: 4, pp. 684–693, Apr 2021.

[8] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin and Nicolas Padoy, (May 2016). "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," [Online]. Available: https://arxiv.org/abs/1602.03012

[9] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim and Nassir Navab, (Mar 2020). "TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks," [Online]. Available: https://arxiv.org/abs/2003.10751

[10] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C. W. Fu, and P. A. Heng, "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Transactions on Medical Imaging,* vol. 37, no. 5, pp. 1114–1126, 2018.

[11] Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C. W. Fu, and P. A. Heng, "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Medical Image Analysis*, vol. 59, 2020.

[12] Ian Goodfellow, Yoshua Bengio, (2016). *Deep Learning,* Cambridge, MA : MIT Press.

[13] Shaojie Bai, J. Zico Kolter and Vladlen Koltun, (Mar 2018). "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," [Online]. Available:https://arxiv.org/abs/1803.01271

[14] R. Araki, T. Onishi, T. Hirakawa, T. Yamashita and H. Fujiyoshi, "MT-DSSD: Deconvolutional Single Shot Detector Using Multi Task Learning for Object Detection, Segmentation, and Grasping Detection," *IEEE International Conference on Robotics and Automation (ICRA),* pp. 10487-10493, Sept 2020.