DEEP LEARNING BASED MULTI-LABEL CLASSIFICATION FOR SURGICAL TOOL PRESENCE DETECTION IN LAPAROSCOPIC VIDEOS

*Sheng Wang, Ashwin Raju, Junzhou Huang**

The University of Texas at Arlington, Dept. of Computer Science and Engineering, Arlington, TX, USA

ABSTRACT

Automatic recognition of surgical workflow is an unresolved problem among the community of computer-assisted interventions. Among all the features used for surgical workflow recognition, one important feature is the presence of the surgical tools. Extracting this feature leads to the surgical tool presence detection problem to detect what tools are used at each time in surgery. This paper proposes a deep learning based multi-label classification method for surgical tool presence detection in laparoscopic videos. The proposed method combines two state-of-the-art deep neural networks and uses ensemble learning to solve the tool presence detection problem as a multi-label classification problem. The performance of the proposed method has been evaluated in the surgical tool presence detection challenge held by Modeling and Monitoring of Computer Assisted Interventions workshop. The proposed method shows superior performance compared to other methods and has won the first place of the challenge.

Index Terms— Surgical tool detection, Deep learning, Multi-label classification, Ensemble

1. INTRODUCTION

Automatic recognition of surgical workflow is an unresolved problem among the community of computer-assisted interventions (CAI). Automatic surgical phase recognition is the basis for many functions in operating room of the future [1] such as monitoring the surgical process and providing assistance during surgery. Besides, it satisfies the current demand to automate the labor-intensive indexing of surgical video databases. Among the recent studies for surgical phase recognition, many studies [2] use surgical triplets (the utilized tools, the anatomical structure, and the surgical actions) of each frame in the videos to represent each time in surgery. Among the three features, the feature of utilized tools is critical to get better phase recognition performance as demonstrated in [3]. To extract this feature leads to a significant problem – surgical tools presence detection.

*Corresponding author: jzhuang@uta.edu. This work was partially supported by U.S. NSF IIS-1423056, CMMI-1434401, CNS-1405985 and the NSF CAREER grant IIS-1553687.

Different from traditional surgical tool detection [4] or tracking [5] problems, surgical tools presence detection does not require the awareness of the locations of surgical tools. Instead, surgical tool presence detection only detects what kinds of surgical tools are used in each time step during surgery. This detection problem can be viewed as an image classification problem while it has several difficulties than the traditional classification problem. For instance, it is probable that multiple tools are used at the same time during the surgical process. Thus, one image frame in a video may contain more than one tool; The frequencies of different tools being used vary a lot, so the detection is harder because of the data imbalance; Plus, the videos have higher quality and more frames because of the improvement of the recording devices. It is almost impossible to manually design feature for tool presence detection. That is why more powerful classification method is necessary to overcome such difficulties.

Deep learning [6], as one of the most successful machine learning methods in the decade, allows deep neural networks discovering the representations from raw data for specific tasks such as classification [7] and detection [8]. It has beaten other machine learning methods in many areas especially when the data set is large enough. Supervised learning is the most common form of machine learning which deep learning improves the state-of-the-art of most problems. For multi-label classification problem, recent studies such as [9] and [10] have demonstrated that deep learning can still achieve good performance. Thus, deep learning seems a promising solution for tool presence detection.

In this paper, we proposed a deep learning based multi-label classification method for surgical tool presence detection in laparoscopic videos. The most related work [3] proposes a novel convolutional neural network (CNN) named EndoNet to solve the phase recognition and tool presence detection tasks in a multi-task manner. Different from their work, we only focus on the tool presence detection to get better performance. The proposed method combines two of the state-of-the-art image classification deep neural networks – VGGNet [11] and GoogLeNet [12]. After training each of the two networks, we use average ensembling technique on the models to avoid overfitting since deep networks with complicated structure are quite powerful but easy to overfit the training data. To evaluate our method, we have submitted

978-1-5090-1172-8/17/$31.00 ©2017 IEEE

our method and experimental result to the surgical tool detection challenge on Modeling and Monitoring of Computer Assisted Interventions (M2CAI) and our method has won the first place of the tool detection challenge.

2. METHODS

Our method follows two main steps: training the CNN models – VGGNet and GoogLeNet, then using ensemble learning to combine the results of the models to get the final results. Before giving the details of the two steps, we will describe the surgical tool presence detection as a multi-label classification problem.

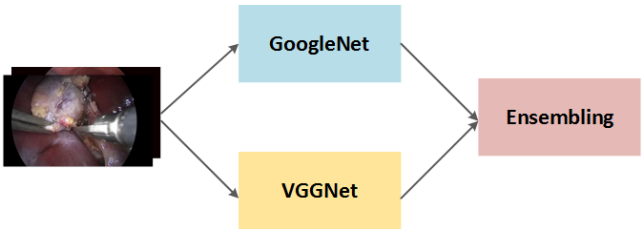

Fig. 1. Pipeline for our tool presence detection method. The left side shows the training image samples, the middle shows two deep neural networks trained from the training images, the right is the ensemble learning technique combining the results of the two models.

2.1. Multi-label Classification

Traditional multi-class classification is the problem of classifying instances into one of the more than two classes, and each instance belongs to only one class. Different from multi-class classification, multi-label classification allows each instance to belong to one or more than one classes. Multi-label classification is a generalization to multi-class classification. In real-world problems, multi-label classification tasks are ubiquitous. For instance, in text categorization, each document can belong to more than one predefined topics, such as sport and health.

The surgical tool presence detection problem can also be viewed as a multi-label classification problem. It is because that each image which we extract as image frames from the surgery videos may contain one or more than one surgical tools. Thus, each image can belong to one or more than one classes. In this way, we can use multi-label classification methods for surgical tool presence detection. The two common methods for multi-label classification are problem transformation and algorithm adaption. Problem transformation decomposes the multi-label classification problem into multiple independent binary classification problems. Algorithm adaptation methods [13] design or adapt algorithms to solve multi-label classification directly. In the proposed method, we use problem transformation method to convert the multi-label classification problem into several independent binary classification problems. Each of the binary classifiers is to detect if one kind of the tools is used in the images.

2.2. VGGNet and GoogLeNet

VGGNet [11]. VGGNet is a deep CNN architecture with 16 layers. Different from other deep CNN architectures, the convolutional layers in VGGNet use very small (3×3) convolution filters. In our training process, we initialize the network weights with the method mentioned in [14]. Rectified Linear units (ReLU) [15] is used as the activation function VGGNet. The batch size used in VGGNet is 32.

GoogLeNet [12]. GoogLeNet is a deep convolutional neural network architecture with 22 layers. GoogLeNet integrates several inception modules inside. The inception modules can increase the depth and width of the network while keeping the computational complexity. GoogLeNet has six more layers than VGGNet but three times fewer parameters compared with VGGNet. GoogLeNet has the ability for multi-scale processing and has achieved state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). In our training process, we use Leak ReLU [16] as the activation function. The batch size used in GoogLeNet is 64.

For both VGGNet and GoogLeNet, we use sigmoid cross-entropy as the loss function and use batch normalization after convolutional layers.

2.3. Model Ensembling

An ensemble [17] consists of a set of independently trained classifiers whose predictions are combined as the final prediction when classifying new instances. Many research studies have shown that good combination of the predictions of multiple classifiers can produce a better classifier.

We use ensembling in our methods for the three following reasons: First, according to the theory of ensemble learning, it is promising to get better classification performance from the ensemble of individually trained classifiers. Second, the process of training deep neural networks tends to overfit the training dataset even if some techniques for avoiding overfittings such as early stopping and Dropout are used in training process or network architecture. Third, data sets provided by challenges always have a larger variance. Thus, even if we get good performance on the validation data set, we cannot assure it will have similar performance on the testing data set.

In the proposed method, we use model averaging to ensemble the predictions from all trained GoogLeNet and VGGNet together to get the final prediction. Simply speaking, we have a prediction probability for each instance from each of the trained models and we calculate the average of the probabilities as the final probability for the instance.

Index	T1	T2	T3	T4	T5	T6	T7
Number	10967	635	14130	411	878	953	1504

Table 1. The numbers of training images for each surgical tool.

3. EXPERIMENTAL RESULTS

To evaluate our method, we have submitted the results of our method to the M2CAI surgical tool presence detection challenge[1] and add some experimental analysis by using the ground truth of the challenge testing data set.

3.1. Data Description and Augmentation

This dataset from M2CAI surgical tool presence detection contains 15 videos of laparoscopic cholecystectomy procedures from University Hospital of Strasbourg/IRCAD (Strasbourg, France). The dataset is split into two parts: the training subset (containing ten videos) and the testing subset (5 videos) by the challenge organizers. In the 15 videos, there are seven kinds of surgical tools in total as shown in Figure 2: grasper, hook, clipper, bipolar, irrigator, scissors and specimen bag. We notate the seven tools from T1 to T7 for short.

Table 2.3 shows the number of training images for each kind of surgical tools in the training set. From the numbers we find that the data set is imbalanced, which makes it more difficult for the models to handle.

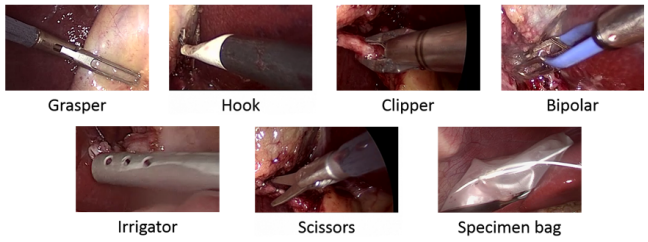

Fig. 2. All the surgical tools in the M2CAI surgical tool detection challenge: grasper, hook, clipper, bipolar, irrigator, scissors and specimen bag.

3.2. Data Preprocessing and Augmentation

Data Preprocessing. We extract the images which have ground truth labels from the ten training videos and resize them into the same size (224×224) since the videos have different dimensions. We use the data from the ten training videos as training and validation sets. For the five testing videos, we extract the images as required by the challenge as the testing set. We also resize them into 224×224.

[1]M2CAI Surgical Tool Presence Detection Challenge 2016: http://camma.u-strasbg.fr/m2cai2016/

Methods	Mean AP
Proposed	**63.8**
Sahu et al.	61.5
Twinanda et al. [3]	52.5
Zia et al.	37.8
Luo et al.	27.9
Letouzey et al.	21.1

Table 2. The leader board of M2CAI surgical tool detection challenge. The evaluation metric is mean accuracy precision.

Data Augmentation. We introduce three kinds of data augmentation methods: horizontal flipping, vertical flipping, and rotation. In the implementation, we do not generate the augmented data set before training. Instead, we dynamically augment each image via each of the three augmentation methods in each epoch of the training process. For each image in a certain training epoch, it has 0.5 probability to be horizontal flipped. It also has 0.5 probability for other two augmentations. The three augmentation methods are taken independently. Thus, we augment our training data set in a dynamic way to better train the models. We do not augment our validation set or testing set.

3.3. Experiment Settings

For both VGGNet and GoogLeNet, we randomly choose 90% data as training set and 10% as testing set five times. Thus, we train ten models in total with five VGGNet models and five GoogLeNet models. We train ten models to let different models have different training data. Then after averaging the ten models, the ensemble will hardly overfit the training data.

3.4. Experimental Results

We use the final prediction ensemble from the ten models on the testing set as the final submission to M2CAI surgical tool presence detection challenge. The mean accuracy precision (mAP) values of all the participants are listed in Table 3.4. The proposed methods have better mAP than the other methods. The method by Sahu et al. has the second best performance by introducing the temporal information to help classification. It demonstrates that our model has excellent performance even not considering the temporal information. Table 3.4 shows the mAPs for each kind of the surgical tools. Our method is still affected by the imbalance of the data set. Further effort should be taken into handling data imbalance.

4. CONCLUSION AND FUTURE WORK

In this paper, we propose a multi-label classification deep learning method for surgical tool presence detection in laparoscopic videos. The proposed method combines two state-of-

Index	T1	T2	T3	T4	T5	T6	T7
mean AP	81.4	62.8	88.2	49.8	49.8	35.3	55.2

Table 3. The mean AP values for each of the seven tool evaluated. These values are computed from our final submission

the-art deep neural networks – VGGNet and GoogLeNet. For better performance, we use ensembling to get the final prediction by averaging all the trained models. The proposed method has better tool presence detection performance than other methods in the M2CAI challenge. However, by directly converting the tool detection to a multi-label classification problem, we have not considered the temporal information of the videos. We will introduce the temporal information to see if it can improve the detection performance in the future.

5. REFERENCES

[1] Kevin Cleary, Ho Young Chung, and Seong K Mun, "Or2020 workshop overview: operating room of the future," in *International Congress Series*. Elsevier, 2004, vol. 1268, pp. 847–852.

[2] Darko Katić, Anna-Laura Wekerle, Fabian Gärtner, Hannes Kenngott, Beat Peter Müller-Stich, Rüdiger Dillmann, and Stefanie Speidel, "Knowledge-driven formalization of laparoscopic surgeries for rule-based intraoperative context-aware assistance," in *International Conference on Information Processing in Computer-Assisted Interventions*. Springer, 2014, pp. 158–167.

[3] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy, "Endonet: A deep architecture for recognition tasks on laparoscopic videos," *arXiv preprint arXiv:1602.03012*, 2016.

[4] Raphael Sznitman, Carlos Becker, and Pascal Fua, "Fast part-based classification for instrument detection in minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 692–699.

[5] Yeqing Li, Chen Chen, Xiaolei Huang, and Junzhou Huang, "Instrument tracking via online learning in retinal microsurgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 464–471.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[9] Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes, "Meka: a multi-label/multi-target extension to weka," *Journal of Machine Learning Research*, vol. 17, no. 21, pp. 1–5, 2016.

[10] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[11] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[13] Min-Ling Zhang and Zhi-Hua Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 10, pp. 1338–1351, 2006.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034.

[15] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[16] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[17] David Opitz and Richard Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.