ISBN 978-981-14-1684-2
Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering
(WCSE 2019 SUMMER)
Hong Kong, 15-17 June, 2019, pp. 427-431
doi: 10.18178/wcse.2019.06.064

FSNet: Pose Estimation of Endoscopic Surgical Tools Using Feature Stacked Network

Yakui Chu [1], Xilin Yang [1], Yuan Ding [1], Danni Ai [1], Jingfan Fan [1], Xu Li [1], Yongtian Wang [1, 2] and Jian Yang [1 +]

[1] Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China

[2] AICFVE of Beijing Film Academy, 4, Xitucheng Rd, Haidian, Beijing 100088, China

Abstract. Identification of surgical instruments is important to understand surgical scenarios and provide assistant processing in endoscopic image-guided surgery. In this paper, we propose a novel feature stacked network (FSNet) for the recognition of surgical tools in endoscopic images. With a lateral connection and concatenation operation on the different layers of the feature pyramid network, high-level semantic information is fused to low-level features, and the bounding boxes are regressed for the tool instance proposals. Then, low-level semantic information is propagated to a high-level network through the bottom-up feature concatenating path. The keypoints of tools are detected in each proposed boundary box. Two state-of-the-art end-to-end tool keypoint recognition networks and three backbones are implemented for comparison. The AP and AR of the our FSNet based on ResNeXt101 are 46.1% and 36.5%, respectively, which surpass the results of other methods.

Keywords: pose estimation, endoscopic image, convolutional neural networks, image-guided surgery

1. Introduction

Image-guided endoscopic surgery has been widely used because of its rapid recovery and less postoperative complications [1]–[3]. However, the endoscopic scenes exhibit different shortcomings, such as visual occlusions, specular reflections and shadows [4], which can deceive surgeons and cause accidental damage in organs and tissues. Precise pose estimation of surgical instruments can provide surgeons with essential information about complex surgical scenarios. It is crucial for improving the accuracy and safety of endoscopic surgery.

Earlier surgical tool detection methods mostly extract low-level visual features [5], [6], such as color, gradient, and texture. However, these methods are inefficient and error prone that small or large tools may be overlooked. With the development of convolutional neural networks, it provides a new approach for surgical tool detection, object segmentation, and pose estimation. EndoNet has successfully shown that CNNs can perform instrument detection and surgical phase recognition [7]. Sarikaya et al. used a multimodal two stream convolutional network based on Faster R-CNN [8] to detect and localize instrument. Du et al. regarded the tools as a combination of several parts connected by joints and used an FCN followed by a regression network and a multi-instrument parsing section [9]. All of these works proved that neural networks can outperform traditional methods in pose estimation and can easily adapt to various surgical conditions.

Low-level features have rich texture details, and high-level features have accurate object position and type information. Thus, concatenating different levels of features can realize the semantic information fusion

+ Corresponding author. Tel.: +010-68913590; fax: +010-68911273.
E-mail address: jyang@bit.edu.cn

of different levels and provide accurate features for the detection of keypoints. In this paper, we propose a novel feature stacked network (FSNet) for the tool pose estimation of endoscopic images. The accuracy of detection and localization of keypoints are improved through the semantic fusion of low-level texture information and high-level position information by connecting inter-level features.

2. Method

The architecture of our proposed feature stacked network for surgical tool pose estimation in endoscopic images is shown in Fig. 1. The first phase is a top-down feature fusion path for tool detection. The output C_i of the backbone network is linked with the corresponding layer P_i of the FPN network through lateral connection, which merges high-level semantic information with low-level features. In the bounding box regression, a unique mask that contains all instances generated for the *M* classes of tools. The second stage is a subnetwork of tool pose estimation, which further fuses the semantic information of different levels through a bottom-up feature cascade path. This path simultaneously passes the feature of the first stage P_i to the final feature map N_i with a lateral connection. Keypoint identification is performed in each proposed boundary box of tool instance, and a unique one-hot mask is generated for each of the *K* keypoints. We connect the detected keypoints according to the predefined skeleton to indicate the pose of the tool.

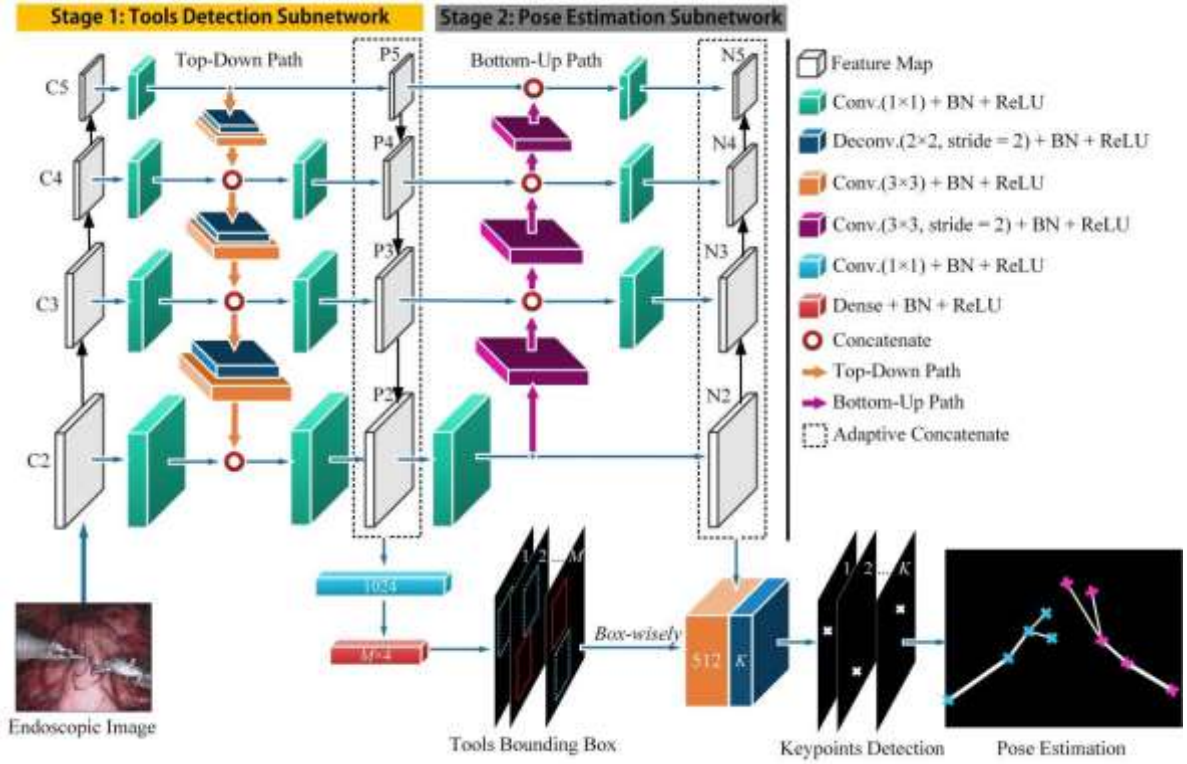


Fig. 1: The architecture of our feature stacked network for endoscopic instrument pose estimation.

2.1. Tools detection and bounding box regression

The top-down feature fusion path is constructed by laterally connecting multi-layers of the backbone. The top layer P_5 of the top-down path is generated from the corresponding output C_5 with a 1×1 convolutional. A 2×2 deconvolutional layer with a stride of 2 is applied to P_5, which outputs a feature map that has the same size as C_4. Similarly, layers $\{P_2, P_3, P_4, P_5\}$ of the top-down path can be acquired.

We perform an adaptive feature connection on each layer of the top-down path, which retains the multi-layer semantic information for the regression of the bounding box. After 1024 1×1 convolution, the bounding box is regressed with a $M\times 4$-kernel fully connected layer, where *M* is the type number of tools. Assuming that t_i is a vector that represents the four coordinates of the predicted bounding box, t_i^* is the true bounding box associated with the positive anchor. The boundary regression loss is defined as

$$L_{reg}(t_i, t_i^*) = R(t_i - t_i^*), \tag{1}$$

where R is the robust loss function ($smooth_{L_1}$) defined in [10].

2.2. Feature stacked pose estimation

The bottom-up path starts from the lowest level N_2, and gradually approaches N_5 with a spatial size downsampling factor 2, as $\{N_2, N_3, N_4, N_5\}$. Each feature map N_{i+1} is generated from a lateral connected feature P_{i+1} and a bottom-up feature N_i. The bottom-up input N_i is dowsampled from higher resolution feature map through a 3×3 convolutional layer with stride 2.

We model the location of a keypoint as a one-hot mask and predict one mask for the *K* keypoint classes (*e.g.*, left clasper, head). Then, for each visible ground-truth keypoint k_i^*, we minimize the cross-entropy loss over an n^2-way *softmax* which output the predicted position of k_i. The keypoint detection loss function is:

$$L_{kp}(k_i, k_i^*) = -k_i \log k_i^* . \tag{2}$$

Thus, we define a loss L for joint training as

$$L(t_i, k_i) = \frac{1}{N_{reg}} \sum_i p_i^* L_{reg} + \frac{1}{N_{kp}} \sum_i L_{kp} , \tag{3}$$

where N_{reg} and N_{kp} are the loss function normalizing parameters of each task, which indicate the number of predicts to be true.

3. Experiments and Results

To evaluate the accuracy of the tool pose estimation, we reimplemented and compared two other state-of-the-art end-to-end networks, Mask R-CNN [11] and PANet [12]. Moreover, we evaluated the impact of two different backbone networks on FSNet, which are VGG19 [13] and Inception V4 [14].

3.1. Experimental setup

Our dataset is acquired from the MICCAI 2017 Endoscopic Vision Challenge[1], which contains six types of surgical tool and 1800 frames of surgical videos. We manually label the visible keypoints of each surgical tool, which contains five keypoints: a handle end, a rigid shaft, an articulated head, the left and right claspers. The clasper points are at the top of the instruments, whereas the head and shaft are localized on the joints.

Our FSNet is based on the ResNeXt101 backbone, and its hyper-parameters are set in accordance with existing PANet works. An anchor is considered positive if it has intersection over union (IoU) with a ground-truth box of at least 0.5 and negative otherwise. Each mini-batch has 1 image per GPU and each image has 512 sampled anchors, with a ratio of 1:3 of positive to negatives. Weight decay is 0:0001, and momentum is set to 0.9.

3.2. Results and discussion

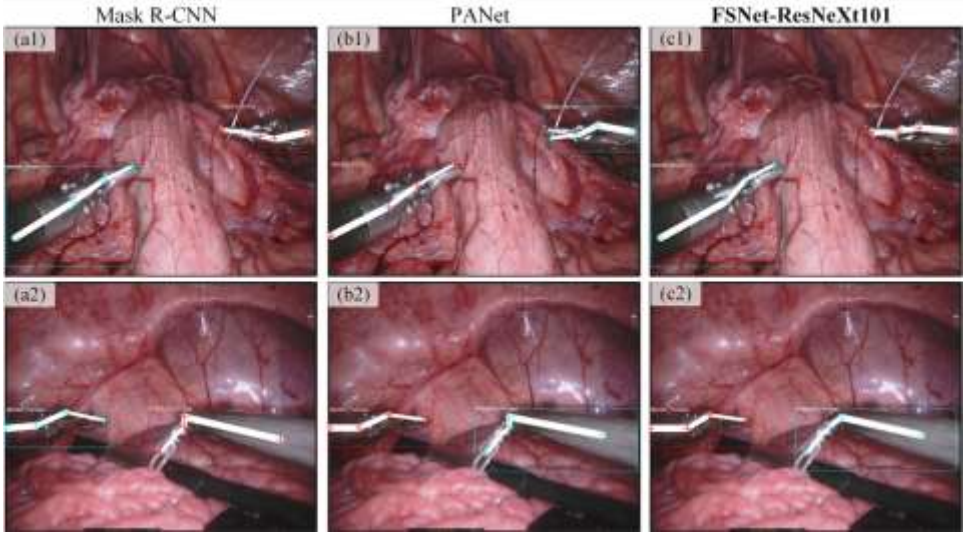


Fig. 2: Results of the comparing methods. Each tool instance is boxed out with a rectangle, and their pose are indicated with a connected line.

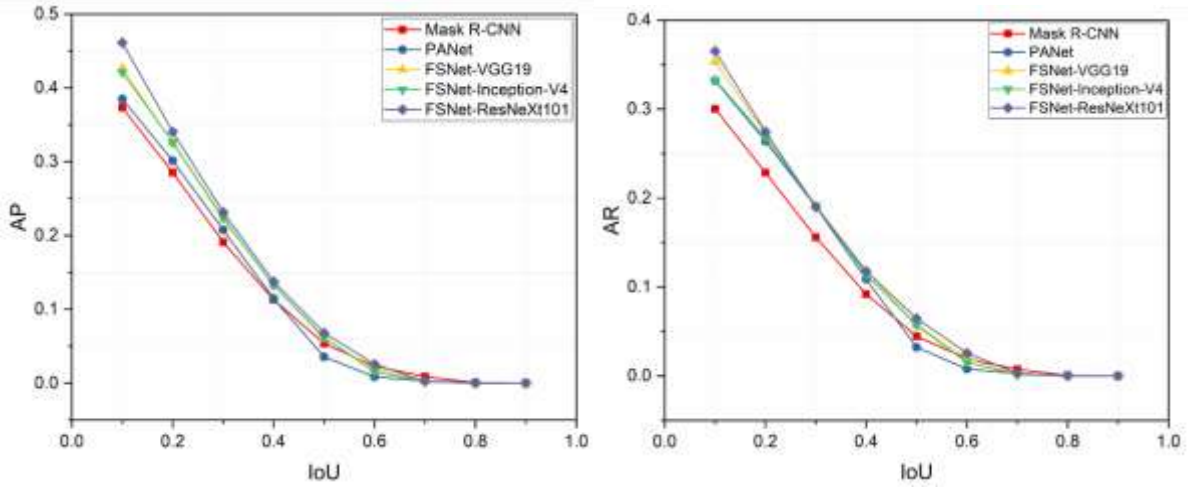
[1] https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/

In the keypoint detection and pose estimation testing experiments, we detect the tool keypoints and connect them according to the skeleton. Each part of the tools is labelled with white line segments of different widths, as shown in Fig. 2. The bounding boxes and names of every detected tool instance are plotted on the images. As shown in the figure, the handle features are obvious and are the easiest to locate and identify, and all three methods perform well. However, Mask R-CNN and PANet methods have errors in the detection of the head and the clasper. Problems such as keypoints lost and mis-positioning also appear. Our method is better than the two other methods in detecting the keypoints, such as handle, rigid shaft, and articulated head, of the tool.

Table 1: Keypoints detection APs (%) and ARs (%) with different radius (pixel).

Method	*backbone*	IoU=0.1		IoU=0.3		IoU=0.5	
		AP	AR	AP	AR	AP	AR
Mask R-CNN	ResNet101	37.3	30.0	19.1	15.6	5.4	4.5
PANet	ResNet101	38.5	33.2	20.8	19.0	3.6	3.2
FSNet	VGG19	42.7	35.3	22.2	19.2	6.1	5.8
FSNet	Inception V4	42.1	33.3	22.5	**19.2**	6.2	5.7
FSNet	ResNeXt101	**46.1**	**36.5**	**23.2**	19.0	**6.8**	**6.5**

In the quantitative evaluation of the keypoints, we create a circular mask with a pixel radius r at each detected keypoint position and its ground truth position. Then, the APs and ARs of different IoUs can be acquired by calculating the overlapping region between the detection keypoint mask and its ground truth. Similar to evaluations of keypoint detection [9], the different radius thresholds indicate the pixel distance between the predicted keypoint and its ground truth. We set the radius threshold r increasing from 5 to 40 and calculate the AP and AR values of different IoUs as radius increasing every 5 pixels. Compared with the original endoscopic image size of 1280×1024, the distance between the detection keypoint and the ground-truth position is about 2.44%–7.81% of the image. The AP and AR values in Table 1 are the mean values calculated with IoU = 0.1, 0.3, and 0.5.



(a) APs at different IoUs (b) ARs at different IoUs

Fig. 3: The APs and ARs of the compared methods.

As shown in Table 1, the proposed FSNet always achieves the highest APs and ARs with different IoUs, which surpasses the other four comparing methods. Moreover, our proposed network surpasses the other two end-to-end networks with three different backbone networks, which proves the efficiency of FSNet. Furthermore, we adopt a visual analysis on the APs and ARs of keypoint detection, as shown in Fig. 3. The keypoint detection APs with different IoUs of five methods are plotted in Fig. 3(a). Obviously, the APs of FSNet are higher than those of other methods. This result proves that the feature stacked concatenating path can effectively improve the accuracy of keypoint detection.

4. Conclusion

This paper proposes a novel FSNet for the recognition of surgical tools in endoscopic images. With the semantic fusion of low-level texture information and high-level position information by connecting inter-level features, the accuracy of detection and localization of keypoints are improved. Two state-of-the-art end-to-end tool keypoint recognition networks and three different backbones are implemented for comparison. The AP and AR of our FSNet based on ResNeXt101 are 46.1% and 36.5%, respectively, which surpass the result of other methods. In addition, the effectiveness and robustness of the proposed method in complex endoscopic scenarios are proven by comparing end-to-end methods and backbone models.

5. Acknowledgements

This work was supported by the National Key Research and Development Program of China (2017YFC0107800), and the National Science Foundation Program of China (61672099, 61527827).

6. References

[1] L. C. Garcia-Peraza-Herrera *et al.*, "ToolNet: Holistically-nested real-time segmentation of robotic surgical tools," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2017–Septe, pp. 5717–5722, 2017.

[2] Y. Chu *et al.*, "Registration and fusion quantification of augmented reality based nasal endoscopic surgery," *Medical Image Analysis*, vol. 42, pp. 241–256, 2017.

[3] Y. Chu *et al.*, "Perception enhancement using importance-driven hybrid rendering for augmented reality based endoscopic surgical navigation," *Biomed. Opt. Express*, vol. 9, no. 11, pp. 5205–5226, 2018.

[4] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: a review of the literature," *Medical Image Analysis*, vol. 35, pp. 633–654, 2017.

[5] N. Rieke *et al.*, "Real-time online adaption for robust instrument tracking and pose estimation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016.

[6] R. Sznitman, C. Becker, and P. Fua, "Fast part-based classification for instrument detection in minimally invasive surgery," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014.

[7] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos," *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 86–97, 2017.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[9] X. Du *et al.*, "Articulated Multi-Instrument 2-D Pose Estimation Using Fully Convolutional Networks," vol. 37, no. 5, pp. 1276–1287, 2018.

[10] R. Girshick, "Fast R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2015 Inter, pp. 1440–1448, 2015.

[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017–Octob, pp. 2980–2988, 2017.

[12] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

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

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.