Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade

Zijian Zhao[1] ✉, *Tongbiao Cai*[1], *Faliang Chang*[1], *Xiaolin Cheng*[2]

[1]*School of Control Science and Engineering, Jinan, Shandong, People's Republic of China*
[2]*Laboratory of Laparoscopic Technique and Engineering, Qilu Hospital of Shandong University, Jinan, Shandong, People's Republic of China*
✉ *E-mail: zhaozijian@sdu.edu.cn*

Published in Healthcare Technology Letters; Received on 24th September 2019; Accepted on 2nd October 2019

Surgical instrument detection in robot-assisted surgery videos is an import vision component for these systems. Most of the current deep learning methods focus on single-tool detection and suffer from low detection speed. To address this, the authors propose a novel frame-by-frame detection method using a cascading convolutional neural network (CNN) which consists of two different CNNs for real-time multi-tool detection. An hourglass network and a modified visual geometry group (VGG) network are applied to jointly predict the localisation. The former CNN outputs detection heatmaps representing the location of tool tip areas, and the latter performs bounding-box regression for tool tip areas on these heatmaps stacked with input RGB image frames. The authors' method is tested on the publicly available *EndoVis Challenge* dataset and the *ATLAS Dione* dataset. The experimental results show that their method achieves better performance than mainstream detection methods in terms of detection accuracy and speed.

1. Introduction: Robot-assisted minimally invasive surgery (RMIS) systems, like the daVinci surgical system (dVSS), have gained more and more attention in recent years. Rather than cutting patients open, RMIS allows surgeons to operate by tele-manipulation of dexterous robotic tools through small incisions, which results in less pain and fast recovery time. With RMIS systems, surgeons sit at a console near the operating table and utilise joysticks to perform complex procedures. Such systems will translate surgeons' hand movements into small movements of the surgical instruments in real time. The location of surgical instruments is a common requirement to provide surgeons with important information for observing tool trajectory and can lighten their burden of finding the instruments during an operation. On the other hand, having real-time knowledge of the motions of the surgical tools can help in the modelling of gestures and skills for the real-time automated surgical video analysis [1], which is good for training the novice surgeons [2]. Hence, in this study, we focus on real-time instrument detection.

Many methods for tool detection have been proposed in the last decade, including optical tracking [3], kinematic template matching [4], image-based detection methods [5], and so on. Nowadays, the image-based (vision-based) methods have become increasingly popular as they require no modification to surgical tool design for providing localisation information [6]. Some early image-based methods utilised low-level feature representations computed over video frames for tool detection, and are comparatively fast. For example, colour segmentation methods [7] by CIELab colour space transformation and thresholding were proposed to extract tool shapes from image frames. Another example of feature representation is gradient features [8] which are often leveraged to retrieve tool edge lines via the Hough transform. However, these methods have significant shortcomings. Noise in image frames, such as lighting change, can easily lead to bad detection results. To overcome these challenges, more robust feature representations, such as scale invariant feature transformation (SIFT) [9] and colour-SIFT [10], have been utilised to detect instruments. Recently, convolutional neural network (CNN)-based methods have become a popular choice for different visual detection tasks such as pedestrian detection, human pose estimation etc. These have also been applied for the analysis of surgical videos, such as instrument presence detection [11, 12], phase recognition [13, 14], tool location [15–17], and tool pose estimation [2, 18]. For example, a cascading model, which consists of a rough location network and a fine-grained search network, was proposed by Mishra *et al.* [19] to locate the tool tip. In the work of Chen *et al.* [20], a CNN is trained with the datasets labelled by a line segment detector to detect a tool's tip, and then the spatial and temporal context algorithm [21] is utilised to detect the tool in real time. These methods exhibit good performance in single surgical tool detection, but they fail to satisfy the need of multi-tool detection. To overcome this problem, a multi-modal CNN based on a faster region convolutional neural network (RCNN) [22] is used by Sarikaya *et al.* [23] for multi-instrument detection. This method achieves good results for detection accuracy but cannot detect the surgical tools in real time (operating at <20 fps).

In this Letter, we propose a novel frame-by-frame real-time detection and location method for multi-instruments, which consists of an hourglass network [24] and a modified VGG-16 [25] network. The former heatmap network is used as a fully convolutional regression network to output the heatmaps which represent the location of the instruments tip area. The latter performs bounding-box regression on these heatmaps stacked with input RGB image frames. In this way, we can simultaneously predict the tools' location and perform recognition. Our method is more like human behaviour. Humans glance at an image and instantly know what objects are in the image and their approximate location (heatmap network), and then locate them precisely in the image (bounding-box network). To evaluate the performance of the proposed method, we evaluated our method on the publicly available multi-instrument *ATLAS Dione* [23] and *EndoVis Challenge* [2] datasets. Our approach obtains better performance than three state-of-art detection methods in terms of detection accuracy and speed.

2. Methodology

2.1. Heatmap network: The overall design of our CNN-based surgical instrument detection model is shown in Fig. 1. This section will describe the architecture of each sub-network of our detection-regression network in more detail. In the proposed framework, the heatmap-regression network (Section 2.1) takes the RGB image frame as input and outputs heatmaps which are

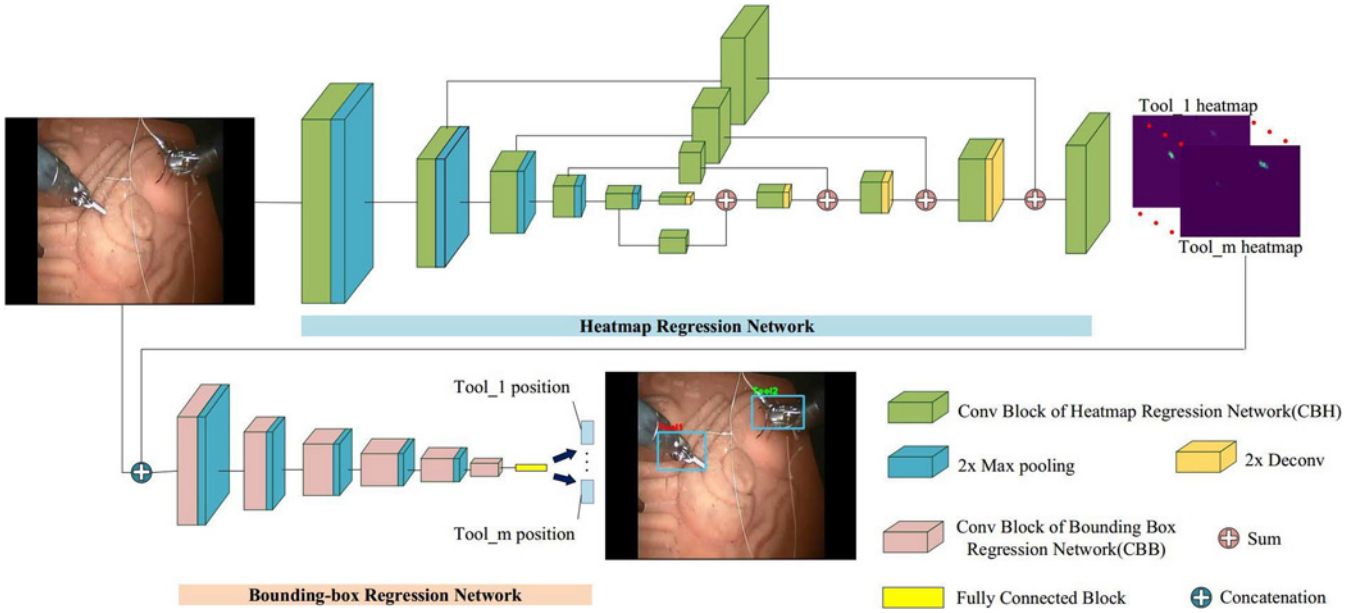

Fig. 1 *Framework of the proposed detection model*

confidence maps of the tool tip areas. These heatmaps guide the bounding-box regression network (Section 2.2) to focus on the location of instruments in the input image. Finally, it outputs four real-valued numbers for each tool which encodes the bounding-box position in the image coordinate system.

In our first sub-network, an hourglass network, which takes an RGB image frame of size $640 \times 480 \times 3$ as input, is employed to output M heatmaps which are confidence maps, one for each surgical instrument. As shown in Fig. 1, the network is composed of 5 maximum pooling layers, 4 upsample layers, and 13 convolutional blocks of which each consists of several residual modules [24]. Thus, the output heatmaps have a resolution of 320×240 pixels. The batch normalisation (BN) layer is added before every rectified linear unit (ReLU) to improve the performance of the network.

We approach a rough instrument location as a binary-classification problem in each heatmap. The ground truth for our regression network is encoded as a set of M binary maps, one for each surgical instrument. As shown in Fig. 2, in each ground-truth heatmap, we set the values within a certain radius around the centre of an object bounding box to 1 as the foreground, and the remaining values are set to 0 as the background. The bigger the object bounding box is, the larger the radius is in the ground-truth heatmap. As we treat the heatmap regression as a multiple binary-classification problem, we train the hourglass network using a pixel-wise sigmoid cross-entropy loss function which is defined as follows:

$$l_{\mathrm{h}} = \frac{1}{M}\sum_{m=1}^{M}\sum_{X(i,j)\in S}[p_{ij}^{m}\cdot\log\hat{p}_{ij}^{m} + (1-p_{ij}^{m})\cdot\log(1-\hat{p}_{ij}^{m})] \quad (1)$$

where p_{ij}^{m} and $\hat{p}_{ij}^{m}$ represent the ground-truth value and corresponding sigmoid output at a pixel location $X(i, j)$ in the mth heatmap of size S.

2.2. Bounding-box network: For the bounding-box regression network, we apply a modified and extended VGG-16 network originally used for image classification, which contains six convolutional blocks, six pooling layers and three fully connected layers as shown in Fig. 1. The BN layer is also added before every ReLU layer. This network takes the heatmaps stacked with the RGB image frame, which is resized to $320 \times 240 \times 3$, as input

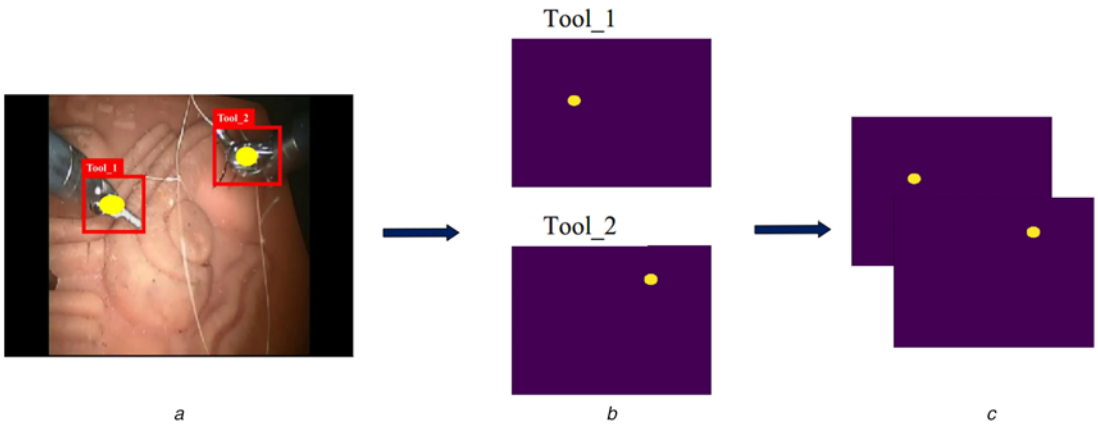

Fig. 2 *Process of obtaining the ground truth of the output of the heatmap network. The yellow coloured area*
a Represents the foreground, and the remaining area is background
b Shows the binary maps, one for each instrument
c Shows the ground-truth heatmaps

doi: 10.1049/htl.2019.0064

and outputs four real-valued numbers that encode the bounding-box position for each instrument in the image coordinate system. The benefit of this stacked architecture is to guide the bounding-box regression network to focus on the instrument tip area.

The goal of this regression network is to predict a precise region box for each instrument. In contrast to bounding-box regression from the faster RCNN, we train our network using a multiple L_1 loss function defined as follows:

$$l_{\mathrm{b}} = \frac{1}{M}\sum_{m=1}^{M} |T_m - \hat{T}_m|. \tag{2}$$

where $T_m(t_m^x, t_m^y, t_m^w, t_m^h)$ and $\hat{T}_m(\hat{t}_m^x, \hat{t}_m^y, \hat{t}_m^w, \hat{t}_m^h)$ represent the corresponding predicted object bounding box and ground-truth bounding box, respectively, in the image coordinate system, and $T(t^x, t^y, t^w, t^h,)$ is defined as

$$t^x = x/W, \quad t^y = y/H, \quad t^w = w/W, \quad t^h = h/H. \tag{3}$$

where x, y, w, h denote the centre coordinates of the box and its width and height. The size of the input image frame is $W \times H \times 3$.

3. Experiments and results: To evaluate the performance of the proposed method, we apply the method to two multi-instrument datasets, namely the *ATLAS Dione* dataset and the *EndoVis Challenge* dataset. We compare the method with three other mainstream detection methods in terms of detection accuracy and speed.

3.1. Datasets and implementation details: The *ATLAS Dione* dataset [11] consists of 99 action video clips of 10 surgeons from the Roswell Park Cancer Institute (RPCI) (Buffalo, NY) performing 6 different surgical tasks (subject study) on the dVSS with annotations of robotic tools per frame. Each frame has a resolution of 854×480. To train our model, we divide the entire set of video clips into two subparts: 90 video clips (20,491 frames) for training and the remaining 9 video clips (1976 frames) for testing. In the MICCAI'15 *EndosVis Challenge* dataset, there are 1083 frames of 720×576 pixels from ex-vivo video sequences of interventions. Similarly, this dataset is separated into a training set (876 frames) and a test set (217 frames). The *ATLAS Dione* dataset is more challenging than the *EndoVis* dataset because there are more disturbing factors, such as motion blurring, fast movement, and background change.

Before training the proposed framework, we initialise both the hourglass network and the modified VGG-16 network using the default initialisation approach in Pytorch 0.4.1. The image frames fed into our model are all resized to 640×480 pixels. We apply a two-step training approach: firstly, we train the hourglass network using stochastic gradient descent (SGD) with a learning rate of 5×10^{-9}, momentum of 0.9, and weight decay of 5×10^{-5}. Then, we keep this fixed and train the modified VGG-16 network using SGD with initial learning rate of 2×10^{-3}, momentum of 0.9, and weight decay of 5×10^{-5}. The learning rate progressively decreases every five epochs by 10%. We implement the proposed detection method and the compared methods (Section 3.2) on Pytorch 0.4.1, Ubuntu 16.04 LTS using an NVIDIA GeForce GTX TITAN X GPU accelerator.

3.2. Comparison with different methods: We compared our method with three other detection methods on the two datasets introduced above. Fig. 3 shows some detection examples in the video frames. In recent years, many object detection models have been applied to surgical instrument detection tasks [1, 23] and achieved great performance. The three anchor-based methods we chose are: Faster R-CNN proposed by Ren *et al.* [22] (the backbone of VGG-16), Yolov3-416 proposed by Redmon *et al.* [26] (the backbone of Darknet-53), and Retinanet proposed by He *et al.* [27] (the backbone of Resnet-50). Non-maximum suppression (NMS) with a threshold of 0.5 is applied to get the final proposals in these methods. Since our proposed anchor-free method does not need extra NMS time, our method has better time efficiency than the other networks. To evaluate the accuracy of our detection method, we use the following evaluation method: if the intersection over union of the predicted bounding box and the ground truth is bigger than 0.5, we consider the instruments to be successfully detected in this frame. As shown in Table 1, our method achieves a mean Average Precision (mAP) of 91.60% for the *ATLAS Dione* dataset, a mAp of 100% for the *Endovis Challenge* dataset and mean computation time of 0.023 s for instrument detection in each image frame. This demonstrates that the proposed method achieves better performance than the other three methods.

We also evaluate our method based on a distance evaluation approach. If the distance between the centre of the predicted

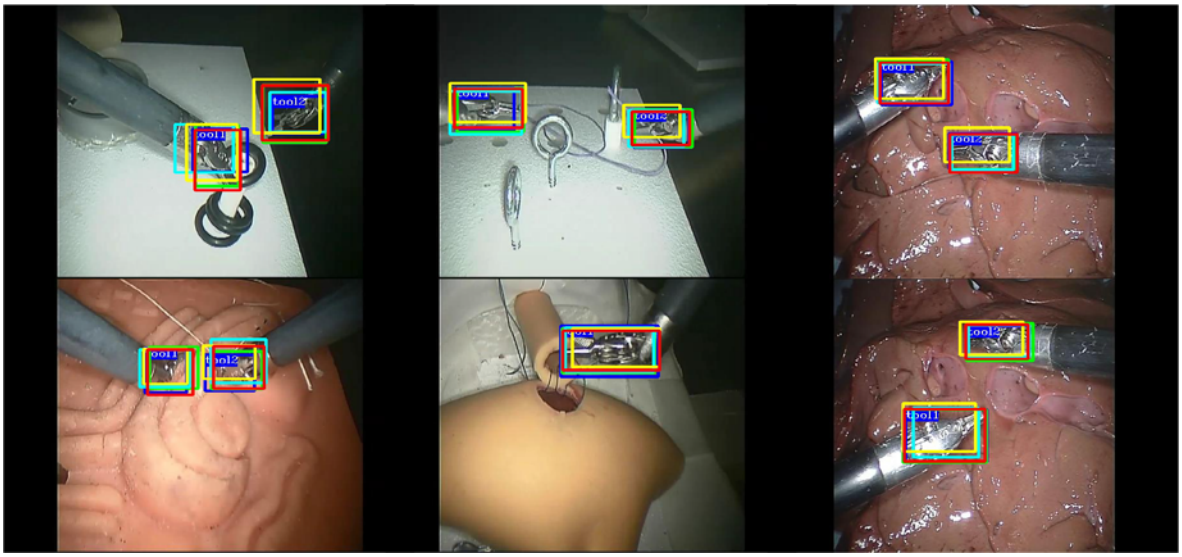

Fig. 3 *Detection examples for two datasets. The two columns on the left are from the ATLAS Dione dataset and the final columns are from the Endovis dataset. As shown in the example frames: our method is in blue, Faster RCNN in green, Yolov3 in cyan, Retinanet in yellow, and the ground truth is in red*

bounding box and the centre of ground-truth bounding-box is less than a threshold in the image coordinates, the surgical instrument is considered to be correctly detected. The experimental results are shown in Fig. 4. Retinanet achieves a better performance than our approach for the *ATLAS Dione* dataset at the cost of 3× lower detection speed. Our method shows the best performance for the *Endovis Challenge* dataset.

Table 1 Detection accuracy and speed of all methods. AP1 and AP2 represent the detection mAP on the *ATLAS Dione* and *EndoVis Challenge* datasets, respectively

Methods	mAP1, %	mAP2, %	Detection time (per frame), s
Faster RCNN (VGG-16)	90.36	100.00	0.064
Yolov3 (Darknet-53)	90.92	99.07	0.034
Retinanet (Resnet-50)	89.39	100.00	0.070
our method	91.60	100.00	0.023

Compared with the other three methods, another advantage of the method is that our method can distinguish between surgical instruments with the same appearance in an image frame. The instruments in two datasets are of the same appearance and the compared method takes them as one class, so they cannot differentiate these instruments. As for our method, the output of the heatmap network is heatmaps which are actually confidence maps, one for each surgical instrument. Fig. 5 shows the RGB input images of which red channel is replaced by the predicted heatmap of each instrument. Based on this, our method can track each instrument although they are of the same appearance in an image frame.

4. Conclusion: In this Letter, we presented a novel frame-by-frame detection method for real-time multi-instrument detection and location using a cascading CNN which consists of an hourglass network and a modified VGG-16 network. The hourglass network is applied to detect a heatmap of each instrument, and the modified VGG is responsible for bounding-box regression. To train our model, we use a two-step training strategy: firstly, we train the hourglass network using a pixel-wise sigmoid cross-entropy loss function, and then keep this fixed and train the

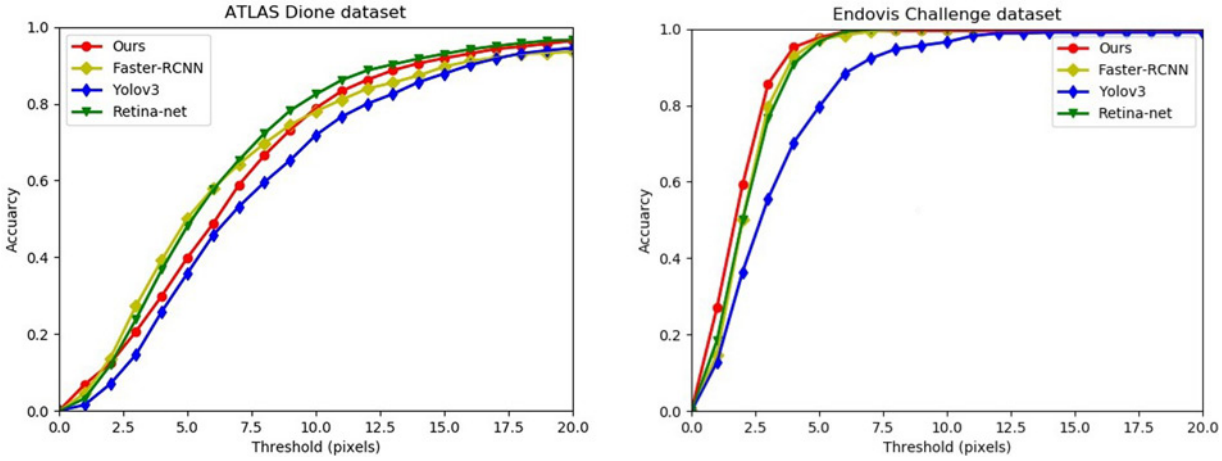

Fig. 4 *Detection accuracy of surgical instrument tips for the two datasets*

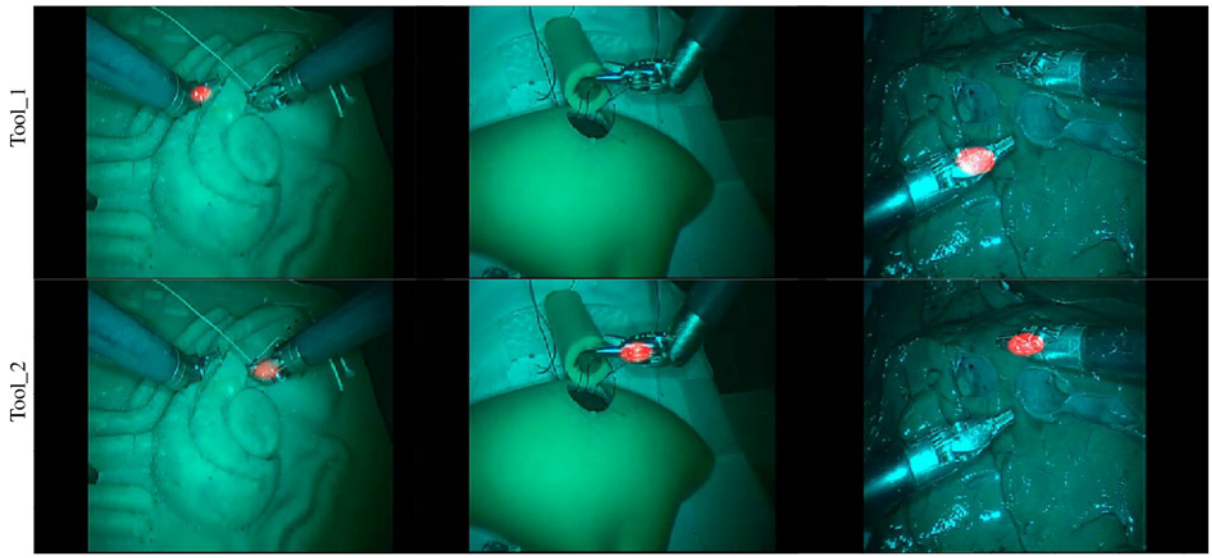

Fig. 5 *RGB input images of which red channel is replaced by the predicted heatmap of each instrument. The two columns on the left are from the ATLAS Dione dataset and the final columns are from the Endovis dataset. The red coloured area represents the tool tip area. As shown in second image of the first rows, Tool_1 is not in this image, so there is no red coloured area, in other words, the values in the heatmap of Tool_1 are close to 0*

doi: 10.1049/htl.2019.0064

whole framework using a multiple L_1 loss function. The proposed detection model is validated on two datasets: the *ATLAS Dione* dataset and the *Endovis Challenge* dataset. The experimental results for these two datasets show that our method achieves a better tradeoff between detection accuracy and speed than the other considered state-of-the-art methods. Moreover, our method can distinguish between instruments of the same appearance while other methods cannot. We think that we can further improve the detection accuracy by replacing VGG-16 with a deeper CNN, but this will reduce the speed correspondingly.

5. Acknowledgments: The authors would like to thank the Qilu Hospital for technical support. This work was supported by the Specialised Research Fund for the Doctoral Program of Higher Education of China (No. 20130131120036), the Promotive Research Fund for Excellent Young and Middle-aged Scientists of Shandong Province (No.BS2013DX027), the National Natural Science Foundation of China (No.81401543, 61273277), the French National Research Agency (ANR) through TecSan Program (DEPORRA).

6 References

[1] Jin A., Yeung S., Jopling J., *ET AL.*: 'Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks'. IEEE Winter Conf. on Applications of Computer Vision (WACV), Lake Tahoe, NV/CA, USA, March 2018, pp. 691–699

[2] Du X., Kurmann T., Chang P.L., *ET AL.*: 'Articulated multi-instrument 2-D pose estimation using fully convolutional networks', *IEEE Trans. Med. Imaging*, 2018, **37**, (5), pp. 1276–1287

[3] Elfring R., de la Fuente M., Radermacher K.: 'Assessment of optical localizer accuracy for computer aided surgery systems', *Comput. Aided Surg.*, 2010, **15**, (1–3), pp. 1–12

[4] Reiter A., Allen P.K., Zhao T.: 'Articulated surgical tool detection using virtually-rendered templates'. Computer Assisted Radiology and Surgery (CARS), Pisa, Italy, June 2012, pp. 1–8

[5] Alsheakhali M., Yigitsoy M., Eslami A., *ET AL.*: 'Surgical tool detection and tracking in retinal microsurgery'. Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling, Orlando, Florida, USA, February 2015, vol. 9415, pp. 11

[6] Bouget D., Allan M., Stoyanov D., *ET AL.*: 'Vision-based and marker-less surgical tool detection and tracking: a review of the literature', *Med. Image Anal.*, 2017, **35**, pp. 633–654

[7] Agustinos A., Voros S.: '2D/3D real-time tracking of surgical instruments based on endoscopic image processing'. Computer-Assisted and Robotic Endoscopy, Springer, Cham, 2015, pp. 90–100

[8] Haase S., Wasza J., Kilgus T., *ET AL.*: 'Laparoscopic instrument localization using a 3-D time-of-flight/RGB endoscope'. 2013 IEEE Workshop on Applications of Computer Vision (WACV), Clearwater Beach, Florida, USA, January 2013, pp. 449–454

[9] Reiter A., Allen P.K., Zhao T.: 'Feature classification for tracking articulated surgical tools'. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Springer, Berlin, Heidelberg, 2012, pp. 592–600

[10] Allan M., Ourselin S., Thompson S., *ET AL.*: 'Toward detection and localization of instruments in minimally invasive surgery', *IEEE Trans. Biomed. Eng.*, 2013, **60**, (4), pp. 1050–1058

[11] Al Hajj H., Lamard M., Conze P.H., *ET AL.*: 'Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks', *Med. Image Anal.*, 2018, **47**, pp. 203–218

[12] Mishra K., Sathish R., Sheet D.: 'Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops, Honolulu, Hawaii, July 2017, pp. 58–65

[13] Twinanda A.P., Shehata S., Mutter D., *ET AL.*: 'Endonet: a deep architecture for recognition tasks on laparoscopic videos', *IEEE Trans. Med. Imaging*, 2016, **36**, (1), pp. 86–97

[14] Zisimopoulos O., Flouty E., Luengo I., *ET AL.*: 'Deepphase: surgical phase recognition in cataracts videos'. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Granada, Spain, September 2018, pp. 265–272

[15] Vardazaryan A, Mutter D, Marescaux J, *ET AL.*: 'Weakly-supervised learning for tool localization in laparoscopic videos'. Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, Springer, Cham, 2018, pp. 169–179

[16] Nwoye C.I., Mutter D., Marescaux J., *ET AL.*: 'Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos', *Int. J. Comput. Assist. Radiol. Surg.*, 2019, **14**, (6), pp. 1059–1067

[17] Laina I., Rieke N., Rupprecht C., *ET AL.*: 'Concurrent segmentation and localization for tracking of surgical instruments'. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Quebec City, Quebec, Canada, September 2017, pp. 664–672

[18] Kurmann T., Neila P.M., Du X., *ET AL.*: 'Simultaneous recognition and pose estimation of instruments in minimally invasive surgery'. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Quebec City, Quebec, Canada, September 2017, pp. 505–513

[19] Mishra K., Sathish R., Sheet D.: 'Tracking of retinal microsurgery tools using late fusion of responses from convolutional neural network over pyramidally decomposed frames'. Int. Conf. on Computer Vision, Graphics, and Image Processing, Springer, Cham, 2016, pp. 358–366

[20] Chen Z., Zhao Z., Cheng X.: 'Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context'. 2017 Chinese Automation Congress (CAC), Jinan, China, October 2017, pp. 2711–2714

[21] Zhang K., Zhang L., Liu Q., *ET AL.*: 'Fast visual tracking via dense spatio-temporal context learning'. European Conf. on Computer Vision, Rich, Switzerland, September 2014, pp. 127–141

[22] Ren S., He K., Girshick R., *ET AL.*: 'Faster R-CNN: towards real-time object detection with region proposal networks'. Advances in Neural Information Processing Systems, Montreal, Canada, December 2015, pp. 91–99

[23] Sarikaya D., Corso J.J., Guru K.A.: 'Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection', *IEEE Trans. Med. Imaging*, 2017, **36**, (7), pp. 1542–1549

[24] Newell A., Yang K., Deng J.: 'Stacked hourglass networks for human pose estimation'. European Conf. on Computer Vision, Amsterdam, The Netherlands, October 2016, pp. 483–499

[25] He K., Zhang X., Ren S., *ET AL.*: 'Deep residual learning for image recognition'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, June 2016, pp. 770–778

[26] Redmon J., Farhadi A.: 'YOLOv3: an incremental improvement', arXiv preprint arXiv:1804.02767

[27] Lin T.Y., Goyal P., Girshick R., *ET AL.*: 'Focal loss for dense object detection'. Proc. of the IEEE Int. Conf. on Computer Vision, Venice, Italy, October 2017, pp. 2980–2988