A Two-Stage Framework for Real-Time Guidewire Endpoint Localization

Rui-Qi Li[1,2], Guibin Bian[1,2], Xiaohu Zhou[1,2], Xiaoliang Xie[1,2], ZhenLiang Ni[1,2], and Zengguang Hou[1,2,3](✉)

[1] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
zengguang.hou@ia.ac.cn

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China

Abstract. The ability of real-time instrument tracking is a stepping stone to various computer-assisted interventions. In this paper, we introduce a two-stage framework for real-time guidewire endpoint localization in fluoroscopy images during the percutaneous coronary intervention. In the first stage, in order to predict all bounding boxes that contain a guidewire, a YOLOv3 detector is applied, and following the detector, a post-processing algorithm is proposed to refine the bounding boxes produced by the detector. In the second stage, an SA-hourglass network modified on stacked hourglass network is proposed, to predict dense heatmap of the guidewire endpoints that may be contained in each bounding box. Although our SA-hourglass network is designed for endpoint localization of guidewire, in fact, we believe the network can be generalized to the keypoint localization task of other surgical instruments. In order to prove our view, SA-hourglass network is trained not only on a guidewire dataset but also a retinal microsurgery dataset, and both achieve the state-of-the-art localization results.

Keywords: Guidewire · Keypoint localization · Surgical instrument

1 Introduction

The keypoint localization of surgical instruments is one of the key components of computer-assisted interventions. From the localization results, we can estimate the pose of the instruments and infer the use status of the instruments. For percutaneous coronary intervention (PCI), the most important surgical instrument is the guidewire which is navigated under real-time fluoroscopy images during the

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32254-0_40) contains supplementary material, which is available to authorized users.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11768, pp. 357–365, 2019.
https://doi.org/10.1007/978-3-030-32254-0_40

intervention, as shown in Fig. 1. Real-time keypoint (i.e. endpoint) localization of guidewire in the fluoroscopy images is of great significance. It can be used in technical skills assessment [1]. More importantly, it could be applied in computer-assisted interventions to help the computer understand the real-time situation.

As far as we know, there is a few research focus on this specific task. Most of research about interventional guidewires focus on guidewire segmentation [2] and the fitting curve [3,4] of the guidewire. Although the endpoint's position of the guidewire can be easily inferred from the segmentation results or the fitting curve results, however, these methods pay more attention to the main body of the guidewire rather than the endpoints. From the results in [2], we can see a median centerline distance error of 0.2 mm but a median endpoint distance error of 0.9 mm. Essentially, the guidewire is a kind of surgical instruments. There has been a lot of research concentrate on the keypoint localization of the surgical instruments used in laparoscopic surgery and retinal surgery [5–7]. Compared with these instruments, the guidewire presents more difficulties so that these methods cannot be applied directly:

1. **Small size of visible part:** Only a small portion of the guidewire is visible, while the main body of the guidewire is almost invisible.
2. **Simple appearance of the endpoint:** Simple appearance seems like an advantage for localization, but it also means there will be more similar structures in the fluoroscopy images, which have a low signal-to-noise ratio.
3. **Non-rigid body:** Not like other surgical instruments, the guidewire is not a rigid body. Therefore, under the premise of a low frame rate (8FPS), the shape of guidewire varies significantly from frame to frame.

To address the above difficulties, a detection stage is proposed before the localization stage inspired by [8]. The overall framework is shown in Fig. 1. In both stages, a method based on deep convolutional neural network (CNN) is proposed. CNN is extremely powerful in extracting local features and performing good predictions utilizing a large receptive field.

Our contributions are as follows. (1) We introduce a cascade framework for guidewire endpoint localization. (2) A post-processing algorithm is proposed in the first stage to deal with the false positives and false negatives of the detections. (3) We also propose a SA-hourglass network in the second stage which can be applied in keypoint localization of other instruments as well. Besides, our framework can achieve real-time localization at an inference rate of approximately 10FPS (fluoroscopy image is about 8FPS).

2 Method

2.1 Stage 1: Guidewire Detection

Our task is to predict a bounding box for each guidewire in consecutive fluoroscopy images. It is different from the detection task which only needs to detect the objects in a single image, also different from the tracking task which needs

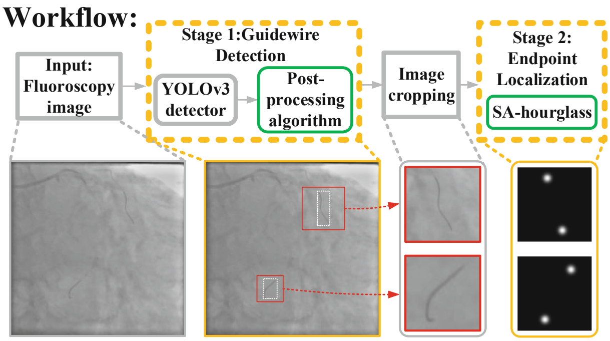
Workflow:
Input: Fluoroscopy image
Stage 1:Guidewire Detection
YOLOv3 detector
Post-processing algorithm
Image cropping
Stage 2: Endpoint Localization
SA-hourglass

Fig. 1. The overall framework for guidewire endpoint localization. In the first stage, detect the location (white box) of all guidewires. Then crop the corresponding patch (red box) from the image. In the second stage, the localization network predicts the heatmaps of two endpoints. The post-processing algorithm and SA-hourglass in green boxes are newly proposed by us. (Color figure online)

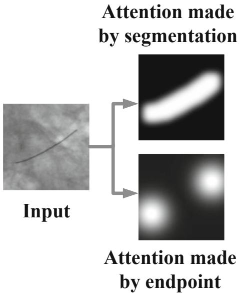

Fig. 2. Demonstrations of two types of attention maps generated by segmentation labels and endpoint positions respectively

to track a class agnostic object. Therefore, a detector can be applied to produce accurate candidates (bounding boxes) of the guidewire, then the constraint relationships between frames can be used to reselect these candidates.

Choosing a Detector: Currently, there are two popular architectures of object detection: one-stage architectures represented by YOLO [9], and two-stage architectures represented by Faster-RCNN [10]. One-stage detectors perform better on speed, while two-stage detectors perform better on accuracy. In order to select an appropriate detector, we train YOLOv3 and Faster-RCNN respectively using our guidewire dataset. Experimental results show that the detection accuracy of YOLOv3 is slightly lower than that of Faster-RCNN (96.4% vs. 98.4% in mAP), but YOLOv3 performs much better than Faster-RCNN in time efficiency (0.05 s vs. 0.12 s). In order to meet the real-time requirement, YOLOv3 is chosen as the detector of our framework. The outputs of the detector are several candidate boxes, each with a confidence score. We only select candidates with scores larger than a given threshold, which is hard to set, as the final outputs.

Post-processing Algorithm: In a continuous sequence of images, there are two primary constraints between two consecutive frames: (1) The distance between the same object in two consecutive frames could not be too far. (2) Existing objects do not suddenly disappear, and objects could not suddenly appear where there was no object before. These two constraints can be used to judge whether the candidate is correct, with the objects existing in the previous frame.

Based on these conditions, a post-processing algorithm is proposed to refine the output candidates of the YOLOv3 detector. Instead of using a single threshold, inspired by the Canny edge detector, all candidates are reselected into two

Algorithm 1. Post-processing algorithm

```
Initialize: $O^t$=$\emptyset$, $O^t_{temp}$=$\emptyset$
Input: $C^t_H$=$\{c_0,...,c_N\}$, $C^t_L$=$\{c_0,...,c_M\}$, $O^{t-1}$, $O^{t-1}_{temp}$
if t==1 then $O^t$=$C^t_H$
else
    for $o_i \in \{O^{t-1}, O^{t-1}_{temp}\}$ do
        $c_{best}$=$c_j$ where $max$(S-IOU$(c_j, o_i)$),$c_j \in \{C^t_H, C^t_L\}$
        if S-IOU$(c_{best}, o_i) \geq \sigma_{IOU}$ then
            add $c_{best}$ to $O^t$; delete $o_i$ from $O^{t-1}$ or $O^{t-1}_{temp}$; delete $c_j$ from $C^t_H$ or $C^t_L$
    for $o_i \in O^{t-1}$ do
        add $o_i$ to $O^t_{temp}$
    for $c_i \in C^t_H$ do
        add $c_i$ to $O^t_{temp}$
return $O^t$,$O^t_{temp}$
```

candidate lists (C_H and C_L) using two thresholds (th_H and th_L, $th_H > th_L$). If the candidate's confidence score is larger than th_H, the candidate is considered highly likely to contain a guidewire and will be put into the list C_H. If the score is less than th_H but larger than th_L, the candidate is considered likely to contain a guidewire and will be put into the list C_L. Two output lists O^t, O^t_{temp} will be created at each timestep t: O^t is used to store the output candidates which already confirmed to contain a guidewire at time t; O^t_{temp} is used to store the temporary output candidates which need to be confirmed in the next timestep. The algorithm is actually to select candidates from two candidate lists C_H and C_L to two output lists O^t and O^t_{temp} at each timestep t with the help of O^{t-1} and O^{t-1}_{temp}. Details can be seen in Algorithm 1. All candidates in O^t and O^t_{temp} are the outputs of the algorithm, also the outputs of the first stage.

Since the shape of the guidewire is variable, a new S-IOU (Intersection over Union) is applied in algorithm: first enlarge each box to a square box by extending the height or width of the box, and then calculate the IOU of two square boxes.

2.2 Stage 2: Guidewire Endpoint Localization

The guidewire endpoint localization component in our framework predicts the heatmaps of two guidewire endpoints, given each bounding box produced by the first stage, as shown in Fig. 1. Because these two endpoints have a similar appearance, we serve both endpoints as the same type of keypoint and predict them in one heatmap. The ground truth of the heatmap is still created by applying a Gaussian kernel to the endpoint's ground truth position as in [7].

Image Cropping: Directly cropping the image by the bounding box and resizing it to the input resolution of the localization network will change the aspect ratio of the guidewire. To keep the aspect ratio of the guidewire, each bounding box is enlarged to a square box by extending either their height or their width.

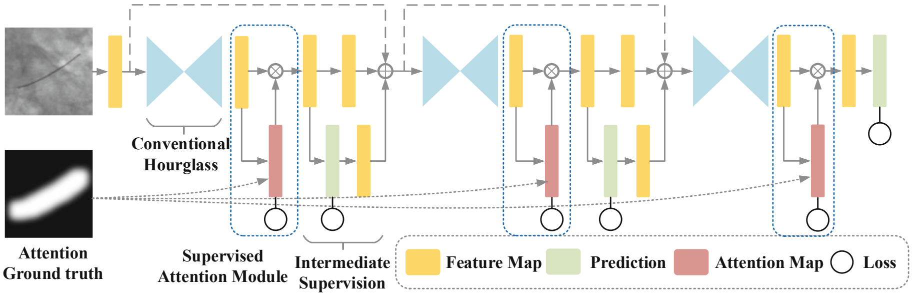

Fig. 3. The proposed SA-hourglass architecture, newly added supervised attention module is shown in the blue boxes. (Color figure online)

The square box is further enlarged with a factor during training and evaluation. During training, a random rescaling factor between 1.1 and 1.3 is applied for data augmentation. During the evaluation, a factor of 1.2 is applied to compensate for possible offsets in the detection results, as shown in Fig. 1.

SA-Hourglass: Stacked hourglass [11] is one of the most popular architectures in human pose estimation. We modify the stacked hourglass by adding a Supervised-Attention (SA) module following the output feature maps of each hourglass and name it as SA-hourglass, as shown in Fig. 3.

Also, some configurations are modified to meet the need of endpoint localization of the guidewire. First, in order to increase the localization accuracy, the first max pooling layer is removed for enlarging the output heatmap size. Second, only three hourglasses are applied in our network. For guidewire and other medical instruments, there is no complex spatial relationships need to learn, so only three hourglasses are applied to reduce the inference time.

Supervised-Attention Module: Our attention module is similar to conventional soft attention in [12]. Following the output feature map of each hourglass, two 3×3 and a 1×1 convolutional layers are applied to generate the attention map. Then the attention map is applied to the feature map which generates it, as shown in Fig. 3. In general, attention mechanism in CNN is used to add a non-linear operation in feature extraction. Since there is no supervision to attention modules, the attention maps learned by the network may not be the results we want. Especially in the heatmap regression, the attention map is supposed to pay more attention around the keypoints, however, because of the pixel-wise distribution of the heatmap is imbalanced, the gradient is dominated by the majority background pixels. As a result, the attention around the keypoints is suppressed, and the focus of attention shifts to the background.

After giving the ground truths to attention maps, SA-hourglass network can be regarded as a multi-task learning network. We propose two methods to generate the ground truth of the attention maps: (1) the same as the ground truth of heatmaps but using a larger Gaussian kernel; (2) additional segmentation labels after several Gaussian filtering. Demonstrations are shown in Fig. 2. Mean-square error (MSE) loss is used in both the attention part and the hourglass part:

$$loss = \frac{1}{wh}(\sum_{x=1}^{w}\sum_{y=1}^{h}(S(x,y) - S^*(x,y))^2 + \lambda\sum_{x=1}^{w}\sum_{y=1}^{h}(A(x,y) - A^*(x,y))^2) \quad (1)$$

In this equation, $S \in \mathbb{R}^{w*h}$ and $A \in \mathbb{R}^{w*h}$ are the predictions of heatmaps and attention maps respectively. $S^* \in \mathbb{R}^{w*h}$ and $A^* \in \mathbb{R}^{w*h}$ are the ground truths of heatmaps and attention maps respectively. λ is for balancing the influence of both loss terms

3 Experimental Results

3.1 Datasets

Two datasets are made to validate our post-processing algorithm and SA-hourglass network respectively. All the images in the two datasets are from in-vivo PCI. And a public dataset is applied to verify the generalization of our SA-hourglass.

Dataset1 consists of 1238 fluoroscopy images with a size of 512*512 (each image contains only one guidewire). All images are randomly divided into the training set (653 images) and the testing set (585 images). We manually label each guidewire's bounding box, segmentation label, and two endpoints' positions.

Dataset2 consists of 10 in-vivo fluoroscopy sequences, with a total of 367 images with a size of 512*512 (contain 609 guidewires in all). Only the bounding box of each guidewire is manually labeled. It should be pointed out that there is no duplicate image between Dataset1 and Dataset2.

The Retinal Microsurgery (RM) dataset [6] contains three video sequences with 1171 images, each with a resolution of 640*480 pixels. Each image contains a single instrument with 4 annotated joints (start shaft, end shaft, left tip and right tip). Analogously to [6], the first 50% of all three sequences is for training and the rest is for testing.

3.2 Implementation Details

For post-processing algorithm, we set σ_{IOU} to 0.3, th_H to 0.3, th_L to 0.01. These two thresholds are obtained through experiments, and they are not difficult to find. We suggest that th_H should not exceed 0.5 and th_L should not exceed 0.1. For SA-hourglass, in data augmentation, random flip, random rotation [$-20°$, $20°$], random grayscale adjustment [-20, 20] and random contrast ratio [0.8, 1.2] are adopted for Dataset1, while only random rotation [$-10°$, $10°$] is adopted for RM dataset. The sigma of Gaussian used in the heatmap's ground truth is 3 for Dataset1 and 7 for RM dataset. λ in loss function is 0.5. The network is implemented using Tensorflow, and for optimization, rmsprop optimizer is applied with a learning rate of 2.5e$-$4 and batch size of 4. Training takes about 13 h on an NVIDIA Titan XP for 500 epochs.

3.3 Detection Experiments

Evaluation Metric: The evaluation metric used in detection experiments is simple: to count the number of true positives (correct), false positives and false negatives (miss) in all frames in test sequences. The correct is defined as the S-IOU score between the detection result and the ground truth exceeds 0.3.

Results: YOLOv3 detector with and without post-processing algorithm are compared in the experiments. The detector has been trained by Dataset1, and Dataset2 is used for evaluation. As shown in Table 1, the results illustrate that the YOLOv3 detector alone works well, but problems remain. And the introduction of our algorithm can significantly reduce the number of false positives and misses in the outputs. From the results, we can also see that it is tough for us to set a single threshold for the detector.

Table 1. Detection results on Dataset2

Detector	Correct	Miss	False positive
YOLO with post-processing ($th_H = 0.3$, $th_L = 0.01$)	604	5	1
YOLO without post-processing (threshold = 0.1)	580	29	26
YOLO without post-processing (threshold = 0.3)	522	87	6
YOLO without post-processing (threshold = 0.01)	608	1	246

3.4 Localization Experiments

Evaluation Metric: Percentage of Correct Keypoints (PCK) metric is used to measure the localization results. PCK reports the percentage of localization results that fall within a distance of the ground truth.

Results: Two state-of-the-art methods on surgical tool keypoint localization are applied for comparison: CSL [7] and in SRPEI [5]. In all, five models are evaluated on both Dataset1 and RM dataset: (1) CSL, (2) SRPEI, (3) Stacked hourglass (3-stack), (4) SA-hourglass with segmentation attention, (5) SA-hourglass with keypoint attention. The results are illustrated in Fig. 4.

From the results, we can see that stacked hourglass's accuracy is significantly improved after SA modules are added (become SA-hourglass). We attribute this improvement to the idea of coarse-to-fine implicitly used in SA-hourglass. Our SA module is designed to generate the coarse attention maps which can eliminate many useless areas of the input. Therefore, we can get more precise results by using a small sigma of Gaussian in the ground truth of output heatmaps. SA module can also be seen as another intermediate supervision with special usage.

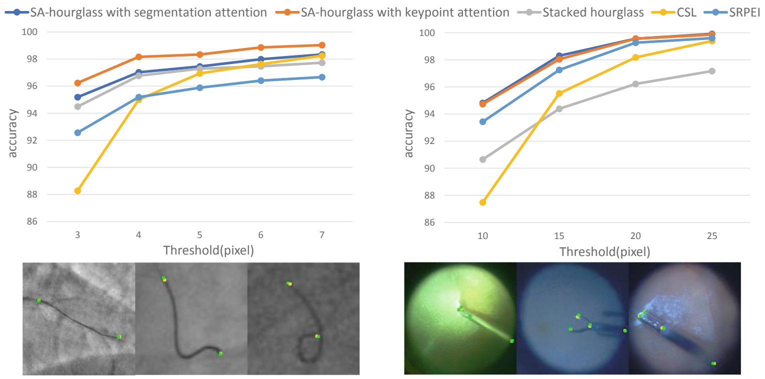

Fig. 4. Average PCK of all keypoints of PCI guidewires (upper left) and RM instruments (upper right). (below) Some localization examples, yellow and green points represent the ground truth and the localization result respectively (Color figure online)

Two kinds of SA-hourglass both achieve the state-of-the-art localization results on both datasets. SA-hourglass with keypoint attention performs best on Dataset1, reaching an accuracy of 96.24% (for threshold = 3), and SA-hourglass with segmentation attention performs best on RM Dataset, reaching an accuracy of 94.82% (for threshold = 10). Besides, the average inference time of our SA-hourglass is about 0.05 s, which fully meets the real-time requirement of fluoroscopy images (8FPS) after adding the detection time.

4 Conclusion

We propose a two-stage framework to localize the guidewire endpoints in real-time fluoroscopy or a fluoroscopy video. For the detection stage, a YOLOv3 detector is applied as a proposal mechanism, and a post-processing algorithm is introduced to refine the bounding boxes produced by the detector. For the localization stage, an SA-hourglass is designed and achieves the state-of-the-art localization results on two datasets. Our framework could be applied to the localization task of other small objects in medical images. As for larger objects, the SA-hourglass network could be directly used without detection stage.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (Grants 61533016, U1713220, U1613210), by the National Key Research and Development Program of China under Grant 2017YFB1302704, by the Strategic Priority Research Program of CAS under Grant XDBS01040100.

References

1. Mazomenos, E.B., et al.: A survey on the current status and future challenges towards objective skills assessment in endovascular surgery. J. Med. Robot. Res. **01**(03), 1640010 (2016)
2. Ambrosini, P., Ruijters, D., Niessen, W.J., Moelker, A., van Walsum, T.: Fully automatic and real-time catheter segmentation in X-Ray fluoroscopy. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 577–585. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_65
3. Vandini, A., Glocker, B., Hamady, M., Yang, G.Z.: Robust guidewire tracking under large deformations combining segment-like features (SEGlets). Med. Image Anal. **38**, 150–164 (2017)
4. Heibel, H., Glocker, B., Groher, M., Pfister, M., Navab, N.: Interventional tool tracking using discrete optimization. IEEE Trans. Med. Imaging **32**(3), 544–555 (2013)
5. Kurmann, T., et al.: Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 505–513. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_57
6. Sznitman, R., Ali, K., Richa, R., Taylor, R.H., Hager, G.D., Fua, P.: Data-driven visual tracking in retinal microsurgery. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012. LNCS, vol. 7511, pp. 568–575. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33418-4_70
7. Laina, I., et al.: Concurrent segmentation and localization for tracking of surgical instruments. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 664–672. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_75
8. Papandreou, G., et al.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4903–4911 (2017)
9. Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
11. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
12. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5669–5678 (2017)