Automatic body region localization in 3D-CT images based on the improved YOLO model

Linghao Du[1], *Rui* Wang[1], *Lin* Cui[1], *Xiaolin* Min[1], *Qingyi* Liu[1], *Yande* Ren[2], *Kai* Huang[2], and *Peirui* Bai[1,*]

[1]College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, Shandong, China
[2] Department of Radiology, Affiliated Hospital of Qingdao University, Qingdao, Shandong, China

Abstract. Automatic body region localization in medical three-dimensional (3D)-CT images is a critical step of computerized body-wide Automatic Anatomy Recognition (AAR) system, which can be applied for radiotherapy planning and interest slices retrieving. Currently, the complex internal structure of human body and time consuming computation are the main challenges for the localization. Therefore, this paper introduces and improves the YOLO-v3 model into the body region localization for these problems. First, seven categories of body regions in a CT volume image I are defined based on the modification version of our previous work. Second, an improved YOLO-v3 model is trained to classify each axial slice into one of the seven categories. Then, the effectiveness of the proposed method is evaluated on 3D-CT images that collected from 220 subjects. The experimental results demonstrate that the slice localizing error is less than 3 NoS (Number of slices), which is competitive to the state-of-the-art methods. Beyond this, our method is simple and computationally efficient owing to its less training time, and the average computational time for localizing a volume CT images is about 3 second, which shows potential for a further application.

Keywords: Body Region Localization, 3D-CT Images, YOLO-v3 Model, Darknet-53.

1 Introduction

1.1 Background and related work

Automatic body region localization based on three-dimensional medical Computed Tomography (3D-CT) is a critical step of AAR, which can be applied for radiotherapy planning and interest slices retrieving etc [1]. The earliest body region localizing method of CT image is to manually extract the scanning and spacing parameters from the header file of DICOM format (Digital Imaging and Communications in Medicine). By this way, the CT

* Corresponding author: bprbjd@163.com

slices corresponding to the interesting body regions or boundaries can be roughly estimated. However, the accuracy and robustness of this method cannot be guaranteed[2,3]. Thereafter, Park et al proposed to establish the mapping relationship between the energy information in the wavelet domain and different imaging parameters of medical images. A body part could be identified and classified by using the lookup table method[4]. Hong et al invented a device and algorithm for locating the neck, thorax, abdomen and pelvis sequentially in terms of the labeled reference frame of the whole body[5]. Both of the localizing methods are applicable to multimodal medical image data. By then, there was no standard definition of the dividing parts and clear expression of the localizing accuracy. Later, an available definition and standard of dividing the human body into five body regions is given by the MIPG in University of Pennsylvania[1,6,7]. A body region was defined by two axial slices: one denoted the superior axial limit or boundary, and the other denoted the inferior axial boundary. Let I represent a scan or image, the location of the superior axial slice of the thorax in I was denoted by $TS(I)$, and the location of its inferior axial slice was denoted by $TI(I)$, as listed in table 1[7]. Similarly, the superior and inferior axial locations of the abdominal and pelvic regions were denoted by $AS(I)$, $AI(I)$, $PS(I)$, and $PI(I)$ respectively. The lower abdominal boundary of human body basically coincided with the upper pelvic boundary. It was noted that $AI(I)=PS(I)$ in terms of anatomical structure. Locations in all images were specified with reference to a fixed scanner coordinate system. The sample slices corresponding to the five detecting region boundaries of the subject were demonstrated in figure 1. It could be observed that each interesting boundary slice had specific visual features (highlighted with green closed lines), which were easily distinguished manually. However, it was difficult to discern them automatically by computer.

Table 1. Definition of body regions and their boundary locations[7].

Body region	Boundaries	Description	Definition
Thorax	*TS*	Thoracic superior axial boundary location	15 mm above the apex of the lungs
	TI	Thoracic inferior axial boundary location	5 mm below the base of the lungs
Abdomen	*AS*	Abdominal superior axial boundary location	Superior-most aspect of the liver
	AI	Abdominal inferior axial boundary location	Point of bifurcation of the abdominal aorta into common iliac arteries
Pelvis	*PS*	Pelvic superior axial boundary location	Inferior boundary of the abdominal region
	PI	Pelvic inferior axial boundary location	Inferior-most aspect of the ischial tuberosities of the pelvis

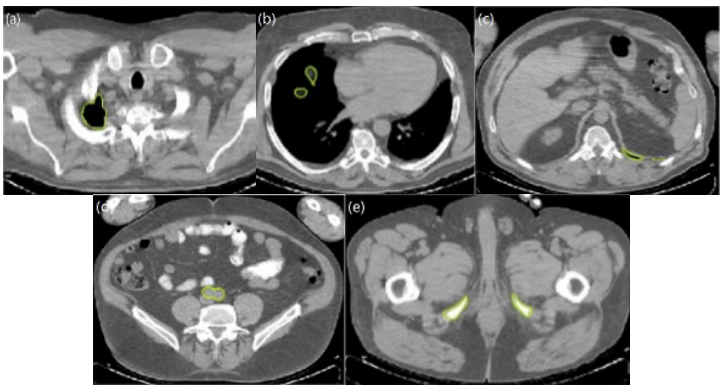

Fig. 1. The demonstration of the five interesting boundary slices of a subject.

The distinguishing features are highlighted in green closed lines. (a) the slice located in 15 mm above *TS*(*I*) which reflects the position of the lung apex, (b) the slice which reflects the position of the superior-most aspect of the liver, (c) the slice located in 5 mm above *TI*(*I*) which reflects the bottom of the lung, (d) the slice which reflects the position of bifurcation of the abdominal aorta into common iliac arteries, (e) the slice which reflects the position of the inferior-most aspect of the ischial tuberosities of the pelvis.

Following the definition of body region mentioned above, there emerged a variety of methods for automatic localizing for human body regions. Hussein et al proposed to train a single image of the upper and lower borders of the abdomen and chest using the convolutional neural networks[8]. However, they only paid attention to the chest and abdomen region which have no high requirement for positioning accuracy. Tong et al proposed a new concept, Virtual Landmarks (VLs), which could be a number of reference points. Then it was adopted to represent the geometric shape of an interesting target[9]. The VLs could not only reduce the computational cost, but also disclose the mapping relationship between multiple targets at different locations. Yan et al proposed to learn the relationship between VLs of parent and offspring objects, the relationship between the VLs of parent object and transformation parameters of child object using neural network, respectively. The trained neural network was then employed to localize and identify the child object[10]. Bai et al proposed to establish the mapping relationship between the reference points of the target and the upper and lower boundary slices of the torso key region by training the neural network. The trained neural network was used to localize the interesting boundary slices simultaneously[7]. Agrawal et al proposed to construct a tandem network structure called BRR-Net by combining the convolutional neural network (CNN) and recurrent neural network (RNN) to localize body regions automatically[11]. The role of CNN is to extract image features and classify each slice of a volume data into one of nine body parts, and the RNN is used to improve the localizing accuracy of low-contrast regions. Although the mentioned methods could achieve acceptable localizing accuracy with a certain degree of complexity and computational cost, it was still difficult to identify the slices in terms of the single characteristics of the organs or tissues due to the complex structures. Otherwise, the traditional machine learning methods were also time-consuming and sensitive to the selection of controlling parameters that usually had strict requirements of the image size and quality.

1.2 Motivation of the work

In this work, a novel body region localizing method was proposed based on the following considerations: 1) The YOLO model was a deep convolutional neural network which can process streaming media video with high computational efficiency. It used a single CNN to detect and classify the interesting objects in real time simultaneously. This was an attractive advantage for dealing with a large number of 3D-CT images (usually contains 200~500 CT slices for a subject). 2) The YOLO model conducted the prediction in terms of the global image information, so it can effectively obtain contextual information and reduce the prediction error. 3) The YOLO model had strong generalization ability, and it was suitable for modifying to train and test the 3D-CT images.

Based on this, the paper adopted the network architecture of the YOLO model as the basic framework. Then experiments on a large database of 3D-CT images were carried out to validate the effectiveness. The experimental results demonstrated that the proposed method is as effective body region localizing approach with competitive performance to the state-of-the-art methods.

The main contributions of this work were as follows: 1) An improved classification network was proposed to automatically locate the body area of 3D-CT images. Since the input CT image did not require to be normalized prior to the training procedure, the useful

information of the original image was preserved well. 2) The classifying accuracy of the network was enhanced by increasing the number of convolutional layers and adopting the feature pyramid network i.e. Darknet-53. The modified architecture could recognize the interesting body slices in a 3D-CT images automatically and efficiently. It showed better robustness as it was insusceptible to the interference factors such as too small targets, complicated target organs or tissues.

The paper was organized as follows. Section 2 introduced the principle and developments of the YOLO model briefly, the modification schemes of the YOLO-v3 network detailly. Section 3 presented the experiments and result analysis. Finally, a conclusion was summarized in Section 4.

2 Methods

The schematic diagram of the proposed method was shown in figure 2. With the purpose of dividing each slice of the input 3D-CT images into one of seven categories, the process was divided into training and testing stages. In the training stage, the category of each slice in the volume data was accurately marked and labeled. Then, these labeled training images were input to the modified YOLO model. In the testing stage, a 3D-CT image in the testing set was input to the trained network, and each slice would be classified into the seven categories i.e. the five body regions, *Legs* and *NOTA*. The *Legs* region refers to the slices below the *PI*(*I*), and the *NOTA* region refers to all the other slices that do not belong to any categories. Then the interesting boundary was be detected and labeled automatically.

2.1 The development of the YOLO model

The YOLO model referred to the regression-based object detector which was an abbreviation of "You Only Look Once"[12]. As mentioned above, it used a single CNN to predict bounding boxes as well as the class probabilities for the boxes simultaneously[13]. More recently, there had been different versions of YOLO-model and their applications, e.g. YOLO-v1 model[12,13,14], YOLO-v2 model[15,16,17], and YOLO-v3 model[18,19] et al. From the point of view of network architecture, the YOLO-v1 model contained 24 convolutional layers and two fully connected layers. The YOLO-v2 model removed the fully connected layers, but added a batch normalization behind each convolutional layer and performs normalization preprocessing for each batch of data. The speed of the algorithm was enhanced further through these modifications. On the basis of the YOLO-v2 model, the YOLO-v3 model added a residual network for every two layers. The network structure was called the Darknet-53. The problem of gradient disappearance or gradient explosion could be alleviated from this architecture modification. Owing to its powerful anti-interference capability and extremely fast computation, the YOLO-v3 model was suitable to detect or classify the interesting objects in real-time.

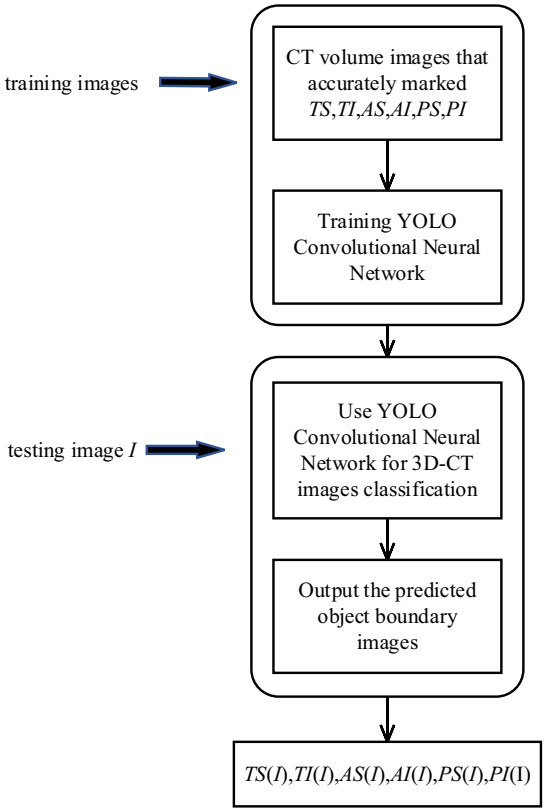

Fig. 2. The schematic diagram of the proposed method.

2.2 The proposed method

In order to train and predict the 3D-CT images by taking advantages of the YOLO-v3 model, we modified the network architecture of the Darknet-53. The modified architecture was shown in figure 3.

The main changes were as follows: 1) The number of the convolutional layers were increased. The original darknet-53 network consisted of 52 convolutional layers and 1 fully connected layer. The last residual part of the network was changed from 4× to 8×, and the number of convolutional layers became 60.

	Type	Filters	Size	output	
	convolutional	32	3×3	256×256	
	convolutional	64	3×3/2	128×128	
1×	convolutional	32	1×1		
	convolutional	64	3×3		
	residual			128×128	
	convolutional	128	3×3/2		
2×	convolutional	64	1×1		
	convolutional	128	3×3		
	residual			64×64	
	convolutional	256	3×3/2	32×32	
8×	convolutional	128	1×1		→scale 1

	convolutional	256	3×3		
	residual			32×32	
	convolutional	512	3×3/2	16×16	
8×	convolutional	256	1×1		
	convolutional	512	3×3		→scale 2
	residual			16×16	
	convolutional	1024	3×3/2	8×8	
8×	convolutional	512	1×1		
	convolutional	1024	3×3		→scale 3
	residual			8×8	

Fig. 3. The modified network architecture based on the YOLO-v3 model.

2) The size of input images was be cropped to 448×448 according to the requirement of YOLO-v3 model. To some degree, this could cause the loss of image characteristics or slowed down the running speed. In contrast, the size of the 3D-CT images was always unchanged as long as the scan parameters keep consistent in our proposed method. Therefore, the CT slices with original size of 512×512 could be directly input to the network for training and testing. That is to say, the proposed method did not need to normalize the image size, which was a time-consuming operation when dealing with large numbers of volumetric images.

3) A smart training trick similar to the fault-tolerant mechanism was adopted to ensure the stable convergence of the modified network in the proposed method. Specifically, the nearest superior and inferior slices adjacent to the true boundary slices were also labeled as the ground truth locations. Therefore, there were three ground truth locations for each detecting region boundary. The final detected slices could be distinguished within the margin of error.

3 Experiments and results

3.1 Data and computational environments

To validate the localizing performance of the proposed method, a 3D-CT image database provided by the Department of Radiology, Affiliated Hospital of Qingdao University was used. The database contained 220 3D-CT images with scanning range from neck to feet (typically comprising of about 300 axial slices). The size of the 2D slices were 512×512, and the average frame spacing was about 5 mm. The volumetric CT images in the databases were dividing into two groups i.e. training set and testing set in terms of a ratio 7:3. For each volumetric CT image I, the true superior and inferior boundaries of the thorax, abdomen, and pelvis was denoted as $TS(I)$, $TI(I)$, $AS(I)$, $AI(I)$, $PS(I)$, and $PI(I)$ respectively. It should be noted that $AI(I) = PS(I)$.

All experiments were performed on a Laptop computer Inspiron 5493 (Dell Inc, The United States) with 128 GB memory, running Python 3.7. The CPU is Intel(R) Xeon(R) Silver 4110 CPU 2.10GHz×2, and the GPU is NVIDIA TITAN XP. The operation system is Ubuntu Server 16.04 64-bit.

3.2 Experimental results and analysis

The localizing results of the proposed method were listed in the last column in table 2. The localizing errors (unit in *N*o*S*) of the existing methods such as the YOLO-v3 model[18], the VL-NN[7], the BRR-Net[11], and the SVM[20] were also presented to make a comparison. It

could be seen that the maximum average localizing errors of the proposed method was 2.7 *NoS* at *AI*, which was obviously less than that of the YOLO-v3 model, i.e. 4.1 *NoS*. The localizing errors of all the five boundaries slices using the two methods confirm that obvious improvement was obtained through the architecture modification of the YOLO-v3 model. In addition, the error values of *TI*(*I*) and *AI*(*I*) are slightly larger than that of the other three boundaries in both the two methods. For the location of *AI*(*I*), the influence of bifurcation of the superior vena cava of the left and right brachiocephalic veins made it difficult to be identified. For the location of *TI*(*I*), it was usually interfered with a black stripe with low image intensity.

The proposed method also demonstrated superior localizing performance than that of the VL-NN and the SVM, except for the localizing error of *TI*(*I*) by SVM method. In contrast, the proposed method showed inferior localizing accuracy compared with the BRR-Net, except for the localizing error of *AI*(*I*). However, the complicated network architectures that containing a tandem architectural CNN and RNN limited the application of BRR-Net, because it needed longer time for training the both network in a two-step strategy. Instead, the average computational time for localizing a volume CT images is about 3 second in our method by taking advantage of the high computational efficiency of the YOLO model. Therefore, the proposed method was competitive with BRR-Net when taking comprehensive considerations of both localizing accuracy and computational complexity.

Table 2. The mean and standard deviation of the localizing errors (unit in NoS) of the five boundaries using the five methods.

Boundary slice	Average localizing errors				
	YOLO-v3[18]	VL-NN[7]	BRR-Net[11]	SVM[20]	Ours
TS	2.3±3.3	2.7±1.8	0.3±0.5	2.3±3.2	1.8±2.7
TI	3.4±4.7	3.0±3.0	1.4±1.7	2.1±1.9	2.4±3.9
AS	2.2±3.5	3.7±1.9	0.6±2.4	2.7±2.4	1.2±3.5
AI=PS	4.1±5.3	3.9±3.2	2.8±2.7	3.3±2.1	2.7±5.2
PI	2.5±3.1	2.5±1.8	0.5±0.5	1.8±2.3	1.3±2.1

4 Concluding remarks

In this work, a novel automatic localizing method of 3D-CT Images was proposed based on the YOLO-v3 model. Five interesting axial body region boundary slices could be detected and localized with satisfactory performance through modification of network architecture and training schemes. Experimental results demonstrated that the proposed method had high computational efficiency and competitive accuracy. In future work, we will continue to increase the size of the dataset and investigate the feature fusion solution of handcrafted characteristics to further improve localizing performance. In addition, the generalization capability of the proposed method for other medical imaging modalities such as MRI, PET, *etc* will be explored.

This work was supported partly by National Natural Science Foundation of China (No.61471225).

References

1. Udupa J K, Odhner D, Zhao L, et al. 2014 Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images Medical Image Analysis vol 18 p 752-771

2. Pianykh O S. What Is DICOM?, In: Digital Imaging and Communications in Medicine (DICOM) 2012 (Springer, Berlin, Heidelberg)
3. Gueld M O, Kohnen M, Keysers D, et al. 2002 Quality of DICOM header information for image categorization Proc. SPIE Medical Imaging vol 4685 p 280-287
4. Park J, Kang G, Pan S, Kim P 2006 A novel algorithm for identification of body parts in medical images. In: Wang L, Jiao L, Shi G., Li X., Liu J. (eds) Fuzzy Systems and Knowledge Discovery. FSKD 2006. Lecture Notes in Computer Science vol 4223 (Springer, Berlin, Heidelberg)
5. L Hong and S Hong 2008 Methods and apparatus for automatic body part identification and localization U.S. Patent App. p 518
6. Wang H, Udupa J K, Odhner D, et al. 2016 Automatic anatomy recognition in whole-body PET/CT images Medical Physics vol 43 p 613-629
7. Bai P, Udupa J K, Tong Y, et al. 2019 Body region localization in whole-body low dose CT images of PET/CT scans using virtual landmarks Medical Physics vol 46 p 1286-1299
8. Hussein S, Green A, Watane A, et al. 2017 Automatic segmentation and quantification of white and brown adipose tissues from PET/CT scans. IEEE Transactions on Medical Imaging vol 36 p 734-744
9. Tong Y, Udupa J K, Odhner D, Bai P, Torigian DA 2017 Virtual landmarks Proc. SPIE vol 10135 p 1013521
10. Yan F, Udupa J K, Tong Y, et al. 2018 Automatic anatomy recognition using neural network learning of object relationships via virtual landmarks Proc. SPIE 10574, Medical Imaging 2018: Image Processing, 105742O, (Houston, Texas, United States)
11. Agrawal V, Udupa J, Tong Y, et al. 2020 BRR-Net: A tandem architectural CNN–RNN for automatic body region localization in CT images Medical Physics vol 47 p 5020-5031
12. Redmon J, Divvala S, Girshick R, et al. 2016 You only look once: unified, real-Time object detection. 2016 IEEE Conference on Computer Vision & Pattern Recognition p 779-788
13. Yali Zheng, Ruikai Zhang, Ruoxi Yu, et al. 2018 Localisation of colorectal polyps by convolutional neural network features learnt from white light and narrow band endoscopic images of multiple databases. Annual International Conference of the IEEE Engineering in Medicine and Biology Society p 4142-4145
14. Parham J, Stewart C 2016 Detecting plains and grevy's zebras in the real world. 2016 IEEE Winter Applications of Computer Vision Workshops. (Lake Placid, USA) IEEE p 1-9
15. Redmon J and Farhadi A 2017 YOLO9000: Better, Faster, Stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR2017) p 6517-6525 (Honolulu, HI, USA)
16. Wang L, Li W, Zhang Y, Wei C 2017 Pedestrian detection based on YOLOv2 with skip structure in underground coal mine 2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC2017) p 1216-1220 (Chongqing, China)
17. Wang M, Liu M, Zhang F, et al. 2018 Fast classification and detection of fish images with YOLOv2. 2018 OCEANS-MTS/IEEE Kobe Techno-Oceans p 1-4 (Kobe, Japan)
18. Redmon J and Farhadi A 2018 YOLOv3: An incremental improvement. arXiv e-prints p 1-6

19. Lin X, Duan P, Zheng Y, et al. 2020 Posting techniques in indoor environments based on deep learning for intelligent building lighting system. IEEE Access vol 8 p 13674-13682
20. Byvatov E and Schneider G 2003 Support vector machine applications in bioinformatics. Appl Bioinformatics vol 2 p 67-77