Automatic detection of prohibited items with small size in X-ray images*

ZHANG Yu-tao (张玉涛)[1], **ZHANG Hai-gang** (张海刚)[2]**, **ZHAO Teng-fei** (赵腾飞)[1], **and YANG Jin-feng** (杨金锋)[2]

1. Tianjin Key Laboratory for Advanced Signal Processing, Civil Aviation University of China, Tianjin 300300, China

2. Shenzhen Polytechnic, Shenzhen 518055, China

(Received 18 July 2019; Revised 3 September 2019)

©Tianjin University of Technology and Springer-Verlag GmbH Germany, part of Springer Nature 2020

In this paper, we focus on the detection of prohibited items with small size, and establish an automatic detection model based on feature fusion single shot multibox detector (FSSD) architecture. Two modifications are carried out to improve the detection accuracy. Firstly, the semantic enrichment module (SEM) with dilated convolution is applied to extract the low level feature with strong semantic information. Secondly, a residual module (Res) with residual blocks is added in the multibox detection architecture in order to extract more adequate features for target detection. The simulation results have demonstrated a better performance of the proposed detection model for prohibited items with small size compared with the state-of-the-arts.

Document code: A **Article ID:** 1673-1905(2020)04-0313-5

DOI https://doi.org/10.1007/s11801-020-9118-x

Security inspection plays a critical role in protecting people from threats. In China, a great deal of transportation requirements have brought enormous work pressure to security inspectors. Taking the civil aviation security inspection as an example, the majority of aviation accidents are caused by human unsafe behaviors. Airport security inspectors as a stress-intensive job, long-term high-stress work environment can cause their work mistakes which affect the safety of aviation operations. It is significant to establish a reliable automatic security inspection system for improving the work efficiency of security inspectors.

X-ray security inspection images are different from natural images and other X-ray images[1]. First, prohibited items in the X-ray security inspection images vary widely in size. Second, the background of the images is messy, which makes it difficult to expect what appears in the background regions. Third, when these items passed an X-ray scan, the penetration property makes it possible to see even the occluded items in the image. Whereas, the occluded prohibited items are not clear. This leads to a difficulty to extract the features of overlapping prohibited items. In summary, it is difficult to detect the prohibited items with small size in X-ray security inspection images.

Improving the quality of feature representations is one of the main technical challenges in small prohibited items detection. In recent years, many researchers have made efforts to further improve the quality of image features on basis of some latest engines, where the most important two groups of methods are feature fusion and learning high-resolution features with large receptive fields[2]. For one thing, as a convolutional neural network (CNN) model consists of a series of convolutional and pooling layers, features in deeper layers will have stronger semantic information. On the contrary, features in shallower layers is not conducive to learning semantics, but it contains more detailed information about edges and contours. Therefore, the integration of deep and shallow features in a CNN model helps improve the quality of feature representations. For another, the small objects occupy fewer pixels in images, the high-resolution features with large receptive fields retain more features of small objects, it is helpful to improve the detection accuracies of small objects.

In this paper, we establish an automatic detection model for prohibited items in X-ray security inspection images. This model adopts fusion single shot multibox detector (FSSD)[3] architecture as the network backbone. In order to detect the prohibited items well in the X-ray security inspection images, a semantic enrichment module (SEM) and a residual module (Res) are added based on FSSD. On the one hand, the SEM takes the low level feature map (the first layer is used to fuse) as input and some dilated convolution layers are applied to generate a semantic meaningful feature map with the same dimension.

* This work has been supported by the National Natural Science Foundation of China (No.61806208), the Fundamental Research Funds for the Central Universities (No.3122018S008), and the Tianjin Education Committee Research Project (No.2018KJ246).

** E-mail: zhg2018@sina.com

The semantic meaningful feature map is used to activate the input layer by element-wise multiplication. The new feature map will replace the original low level feature map for fusion. Dilated convolution is a piratical method to increase both of the receptive field and feature resolution. Its main idea is to expand the convolution filter and use sparse parameters. The semantic information, which is instrumental in small objects detection, is also enriched by increasing receptive field. On the other hand, in order to extract more adequate features and improve the performance of the deep network, an Res is used in this architecture. Res mainly consists of residual blocks and deep convolution layers. The adequate features, which are propitious to detecting small objects, is easy to be extracted by the deep convolution layers. The residual blocks make the deep network maintain good performance of extracting features. The detection model for prohibited items is illustrated in Fig.1.

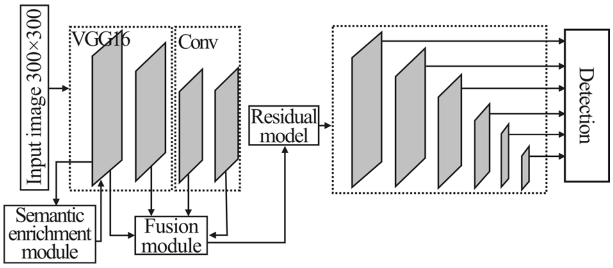

Fig.1 Prohibited items detection net based on FSSD

In the filed of prohibited items detection, a lot of works have been done by the researchers. Mery et al[4] detect the prohibited items in X-ray security inspection images by constructing representative dictionaries. Roomi et al[5] use image segmentation method to extract the region of interest by designing handcrafted features. Turcsany et al[6] use a novel Bag-of-Words representation scheme and speeded-up robust features for image classification and detection. Unfortunately, due to the complexity of X-ray security inspection images, these methods, which manually extract features, cannot detect the prohibited items well.

Object detectors are divided into traditional object detectors[7] and object detectors based on deep learning[8]. Recently, with the rapid development of deep learning, especially the convolutional neural network (CNN), a lot of detectors based on CNN have been proposed in object detection tasks. SSD[9] generates multi-layer feature maps to detect the objects. Deconvolutional single shot detector (DSSD)[10] improves the detection accuracy by adding several deconvolution layers. FSSD[3] concatenates different sizes of feature maps, which come from different layers, and generates feature pyramid to predict detection results directly.

FSSD merges the context information by concatenating feature maps of different sizes. Concatenating different sizes of feature maps is a way to fuse the feature. As can be seen from Tab.1, FSSD is fit for detecting the prohibited items with large size while the accuracies in detecting small size prohibited items need to be improved. In order to solve this problem, an SEM and a Res are added based on FSSD.

Tab.1 The accuracies of different methods

	SSD	FSSD	Ours
mAP	84.3	88.9	91.2
Power bank	90.6	90.8	90.9
Lighter	72.1	87.1	89.2
Fork	80.3	82.7	90.8
Knife	77.4	83.8	86.8
Gun	97.0	98.8	99.1
Scissor	88.5	89.9	90.3

In order to improve the quality of image features and enrich the semantic information of low level feature map, SEM is attached to it. It is a simple network for the SEM. This module mainly composed of dilated convolution layers. In the SEM, the kernel sizes of these dilated convolution layers are 3×3. The first three dilated convolution layers have a dilation rate of 2 and the last dilated convolution layer has a dilation rate of 4.

The proposed procedure of SEM is shown in Fig.2. Firstly, the SEM takes the low level feature map (the first layer is used to fuse) as input. Secondly, four dilated convolution layers are applied to generate the semantic meaningful feature map. Finally, the semantic meaningful feature map is used to activate the input layer by element-wise multiplication. The semantic meaningful feature map and the input low level feature map have the same dimension.

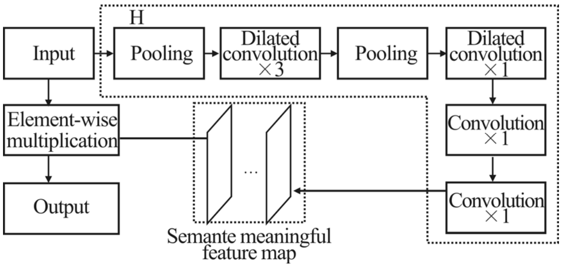

Fig.2 The Semantic enrichment module

Mathematically, let $X\in R^{C\times H\times W}$ be the input feature map, $Y\in R^{C\times H\times W}$ be the semantic meaningful feature map. The X will produce Y by:

$$Y=H(X)\in R^{C\times H\times W}\ , \tag{1}$$

the semantic meaningful feature map Y is used to activate the input feature map X by element-wise multiplication:

$$Z=\ X\odot Y, \tag{2}$$

where Z is the semantically activated low level feature map. Z have both detailed information and semantic information. Z will replace the original X in the feature pyramid for fusion.

In FSSD, after concatenating the feature map, six down-sampling blocks are used to generate new feature pyramid, as shown in Fig.3(a). In order to extract more adequate features and detect the prohibited items with small size well, some additional convolution layers are added based on these down-sampling blocks and the residual blocks, as shown in Fig.3(c), are applied to avoid the degradation problem. The structure of Res is shown in Fig.3(b).

The Res used in this paper is mainly inspired by the residual network[11]. It is an improvement based on FSSD. We add some additional convolution layers and insert shortcut connections to implement Res. The convolution layers used in Res have 3×3 filters and the layers have the same number of filters if they have the same output feature map size. The shortcut connections can be directly used when the input and output are of the same sizes. When the shortcuts go across with two sizes, the number of filters is doubled and the down-sampling operation is carried out for the larger one. The residual operation is performed by element-wise. As shown in Fig.3(c), the output feature map Z is produced by:

$$Z= X\oplus Y, \tag{3}$$

where X is the input feature map and Y is the intermediate feature map. In feature pyramid, each layer of the feature map, which is generated by Res, has the same sizes as the FSSD.

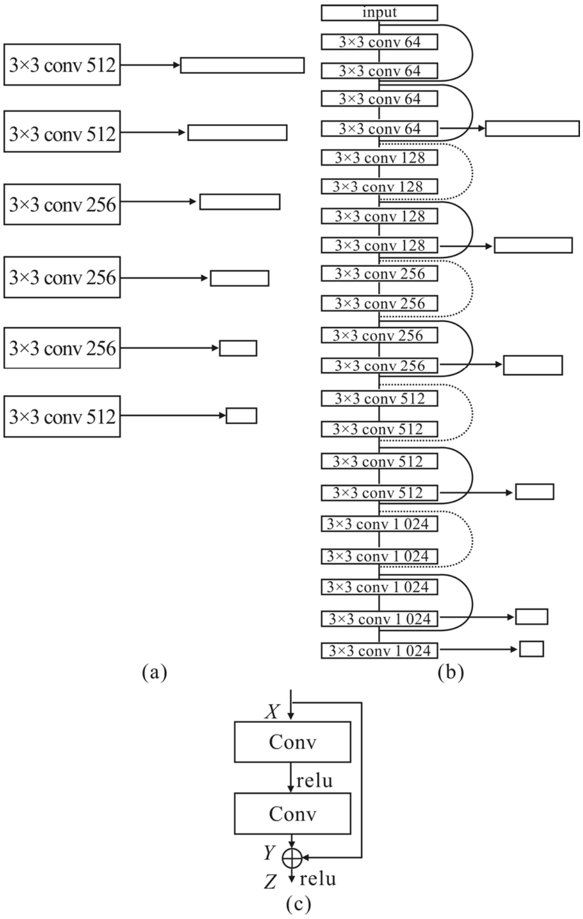

Fig.3 The residual module

In this section, we manifest the effectiveness of the proposed method through several contrast experiments and select the best model by carrying out some ablation studies. The experimental results indicate that the method, which is proposed in this paper, acquires satisfactory performance in terms of detecting the prohibited items with small size. In the following subsections, the database is introduced first, then a series of contrast experiments are carried out, and finally the ablation studies are performed.

Two kinds of databases, which called database A and database B, are used in this paper. There are six common categories of prohibited items, namely, power bank, lighter, fork, knife, gun and scissor in both database A and database B. The database A contains a total of 4 252 X-ray security inspection images. These images are mostly obtained from X-ray scans on personal luggage, in which the size of the objects vary widely and the items are often randomly stacked. These images in the database A are divided into two classes: images with a complex background containing only one category of prohibited item, which are called simple images, and images with a complex background containing two or three categories of prohibited items, which are called complex images. The number of simple images is 2 074 and the number of complex images is 2 178. The size of the images in the database A are 300×300. We randomly divided database A into two subsets for training and testing. There are 2 952 images in the training subsets and 672 images in the testing subsets. All of the images in the testing subsets come from complex images.

In order to enable the detection model learn features of X-ray prohibited items with small size well and avoid the interference caused by the complex background in the X-ray security inspection images when the detection model is trained, database B is made by us. It is only used to train the model. Database B has a total of 1 645 images. These images only contain the foregrounds of prohibited items. The foregrounds of prohibited items are extracted from the collected X-ray security inspection images according to the image preprocessing method. The sizes of the images in the database B are all 300×300. The examples of the database A and the database B are shown in Fig.4. The first line of Fig.4 shows the simple images, the second line of Fig.4 shows the complex images and the third line shows the database B. On the entire database A and database B, we manually add a bounding-box for each prohibited item.

In this paper, we adopt FSSD as the network backbone and carry out two modifications based on it to improve the detection accuracy of prohibited items with small size. In order to evaluate the performance of our utilized model quantitatively, we conduct the contrast experiments on these two databases mentioned above and compared the results of our model with SSD and FSSD. The detection accuracies of different models are shown in Tab.1. The detection results of the three detection models are shown in Fig.5.

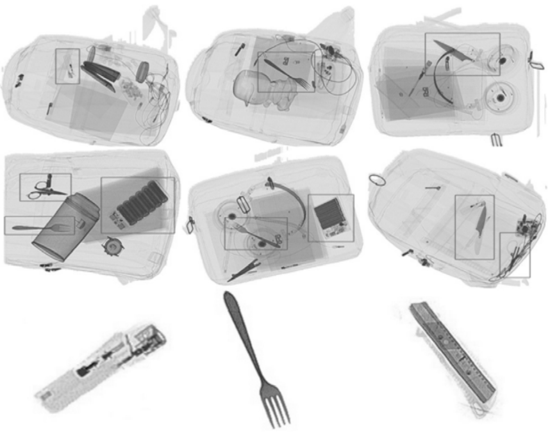

Fig.4 Examples of images in database

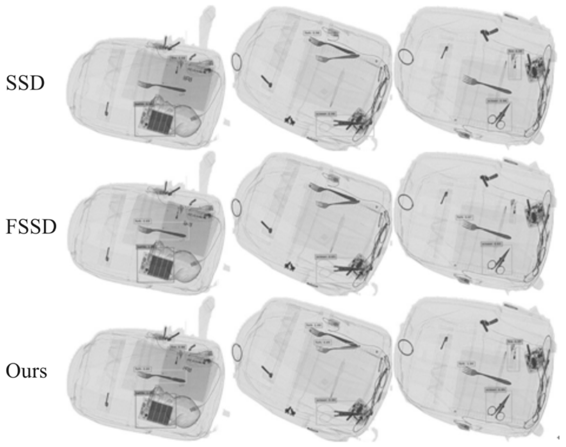

Fig.5 The detection results of different methods

We use the following strategies for training. Firstly, the database B is used to train the model. Secondly, we load the weights obtained by training database B into the network and retrain the model by using training subset of database A. Finally, the testing subset of database A is used to evaluate the performance of the detection model. These models are trained on an Nvidia 1080Ti GPU with batch size 16. The initial learning rate is set to 0.000 1. The weight decay is set to 0.000 5. The predicted bounding box is correct if its intersection over union (IoU) with the ground truth is higher than 0.5. We adopt the mean average precision (mAP) as the metric for evaluating detection performance.

When coupled with Tab.1 information, leads to some possible conclusion that the proposed model obtains a higher accuracy than SSD and FSSD. Compared with SSD and FSSD, mAP increased by 6.9% and 2.3%, respectively. From the third row and the seventh row of Tab.1, we can conclude that by employing SSD and FSSD, the detection accuracies of power bank are 90.6% and 90.8%, the detection accuracies of gun are 97.0% and 98.8%. Both SSD and FSSD are suitable for detecting the prohibited items with large size. Nevertheless, SSD does not obtain satisfactory results of detecting smaller prohibited items, such as lighter, fork and knife. Even if FSSD obtains higher mAP than SSD when detecting the smaller prohibited items mentioned above, our proposed model acquires best classification performance among these three methods. For example, our model improves the detection accuracy of lighter from 72.1% to 89.2% and from 87.1% to 89.2% for SSD and FSSD, respectively. It can be concluded that our model significantly better than SSD and FSSD on detecting the prohibited items with small size.

To further understand the effectiveness of our two modifications, we do experiments with different settings and report the results in Tab.2. These experiments performed the same training strategies as the contrast experiments. As can be seen from Tab.2, the SEM can improve the performance by 1.5%, which shows the effectiveness of this module. With the Res added, the performance can be further improved. Another ablation study conducted is the position of the SEM. To do this, we place SEM in different parts of the network. Firstly, we only add SEM to the low level feature map of the first feature pyramid (SEM+FSSD). Secondly, the SEM is added on the low level feature map of the second feature pyramid (FSSD+SEM). Finally, SEM is added on both of the low level feature map (SEM+FSSD+SEM). Experiments show that SEM+FSSD yields the best performance, 0.5% better than FSSD+SEM and 0.8% better than SEM+FSSD+SEM. That means the low level feature map of the second feature pyramid contains more semantic information after concatenating feature maps. When the SEM attached to it, SEM interfered with the original information of this layer.

Tab.2 Ablation result

Model	mAP
FSSD	88.9
SEM+FSSD	90.4
FSSD+SEM	89.9
SEM+FSSD+SEM	89.6
SEM+FSSD+Res	91.2

In this paper, we adopt the FSSD to detect the prohibited items in X-ray security inspection images and introduce semantic enrichment module and residual module to improve the detection accuracy of small prohibited items. From the result, we can conclude that these two modules are beneficial for improving detection accuracy of prohibited items with small size.

References

[1] Wang X, Peng Y, Lu L, Lu Z, Bagheri M and Summers R M, ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, IEEE Conference on Computer Vision and Pattern

Recognition, 2097 (2017).

[2] Zou Z, Shi Z, Guo Y and Ye J, Object Detection in 20 Years: A Survey, arXiv: 1905.05055v2, (2019).

[3] Li Z and Zhou F, FSSD: Feature Fusion Single Shot Multibox Detector, arXiv: 1712.00960, (2017).

[4] Mery D, Svec E and Arias M, Object Recognition in Baggage Inspection Using Adaptive Sparse Representations of X-ray Images, Image and Video Technology, 709 (2015).

[5] Roomi M and Rajashankarii R, International Journal of Computer Science, Engineering and Information Technology **2**, 187 (2012).

[6] Turcsany D, Mouton A and Breckon T P, Improving Feature-based Object Recognition for X-ray Baggage Security Screening using Primed Visual Words, IEEE International Conference on Industrial Technology, 1140 (2013).

[7] Wang M, Chen J, Gao F and Zhao J, Optoelectronics Letters **14**, 67 (2018).

[8] Joseph Redmon and Ali Farhadi, YOLO9000: Better, Faster, Stronger, Computer Vision and Pattern Recognition, 7263 (2016).

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu and Alexander C, Berg and: SSD: Single Shot MultiBox Detector, European Conference on Computer Vision, 21 (2015).

[10] Fu C Y, Liu W, and Ranga A, Tyagi A and Berg A C, DSSD : Deconvolutional Single Shot Detector, arXiv: 1701.06659, (2017).

[11] He K, Zhang X, Ren S and Sun J, Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition, 770 (2015).