A Multi-view Deep Learning Approach for Detecting Threats on 3D Human Body

Zhicong Yan[1], Shuai Feng[1], Fangqi Li[1], Zhengwu Xu[2], and Shenghong Li[1]

[1] School of Cyber Security,
Shanghai JiaoTong University Shanghai 200240
[2] School of Information and Communication Engineering,
University of Electronic Science and Technology of China 611731
zhicongy@sjtu.edu.cn
feng_shuai@sjtu.edu.cn
solour_lfq@sjtu.edu.cn
zwxu@uestc.edu.cn
shli@sjtu.edu.cn
http://www.springer.com/lncs

Abstract. *Deep Neural Network-based methods have recently shown outstanding performance on object detection tasks in 2D scenarios. But many task in real world require object detection in 3D space. In order to narrow this gap, we investigate the task of detection and localization in 3D human body in this paper, and propose a multi-view based deep learning approach to solve this issue. Experiments show that the proposed approach can effectively detect and locate specific stuff in 3D human body with high accuracy.*

Keywords: Multi-View Convolution Neural Network, 3D Object Detection, Airport Security.

1 Introduction

Nowadays, airport security is an indispensable requirement for safe travel, which consumes a great number of manpower, forces the passengers into waiting in the long lines and wasting time. Quickly and accurately detecting suspicious stuff camouflaged in human body or luggage via machine learning algorithm is one promising solution.

During the past two decades, most countries have attached great importance to the security of transport infrastructures. scanner machines with new technology and programs are invented to enforce protection in the airports. According to the reports of Transportation Security Administration(TSA), tens of thousands of High Definition-Advanced Image Technology(HD-AIT) system had been deployed in USA airports[1], which can generate 3D millimeter wave scan data of about 2 gigabytes per person. In the practical applications in airports, we have to predict whether any suspicious stuff is present in human body and which part of human body contains them. This problem can be solved as a multi-label

binary classification task, while the difficulty with this problem is taking such high dimensional data as input. For 3D model, the dimension of input data is almost three or four orders of magnitude more than the 2D image data, posing great challenges to design the algorithm.

The convolution neural network(CNN) could efficiently classify 2D image data because its locally connected with weight sharing structure and highly non-linear model function[2]. But when it comes to 3D data with much higher dimension, we need to take more consideration. In 3D scenario, the classification task has been investigated by a number of works[3–5]. The state-of-art method MV-CNN [3] applies the multi-view strategy, which reduces the input dimension by taking multiple 2D views of 3D model as input. But when we applying the MV-CNN to the passenger screening task, we find that the network is confused by the high-dimensional 3D data because the stuff we want to detect only takes a small part in 3D image. So we resort to efficient detection methods in 3D scenario.

For object detection and localization in 2D scenarios, He Kaiming et al has proposed Faster R-CNN algorithm[8] which has been used extensively and investigated by various research paper[8–10]. But there is little research in the field of object detection and localization in 3D scenarios.

For the above reasons, in this paper, we propose a network architecture that can detect and locate suspicious stuff in 3D human body, which adopts the multi-view strategy in MV-CNN[3] as well. We incorporate the detection strategy to the network to make it capable of detection and localization in 3D space. Further, in order to train this network, we split the network to two stages. In the first stage, a pre-trained deep CNN is used to extract score maps from images by different views of 3D human body, predicting score maps at each view, which only contain the position information of suspicious stuff. Then in the second stage, we aggregate those maps into a single feature and use a regularized Multilayer Perceptron(MLP) network for detection. This two stage architecture ensure that only the interested objects are detected by the network. The contributions of this paper are as follows:

1. First, we describe a two stage approach that allows the network to perform detection in 3D space.
2. Second, we validate our approach on the Passenger Screening Dataset, and the result shows that the proposed approach can efficiently detect and locate suspicious stuff in 3D human body.

The rest of this paper is organized as follows: In the section II, we provide the related research work. Section III describes the proposed approach in detail. In section IV and section V we provide the experiment results and conclusion of this paper.

2 Previous Work and Literature

Our method is related to prior work on image-based CNNs and object detection methods. Next we discuss representative work in these areas.

2.1 Convolution Neural Networks

Our work has employed the convolution neural networks, which have recently enjoyed a great success in large-scale image and video recognition. In particular, CNNs have been used as an general purpose image feature extractors for lots of tasks, such as object classification, object detection and semantic segmentation[2, 11, 12, 15]. Without handcrafted features, CNN is easy to train with back-propagation algorithm. The Resnet[11] network and VGG[12] network structure are now the most popular CNN structures in image classification task. We choose the VGG16 network and Resnet-101 network as the base network because they offer a balance between computation and network generalization capability.

2.2 Deep Network for Object Detection

With the development in both architecture and train methods of deep CNNs, object detectors like OverFeat[13] and R-CNN[6] showed dramatic improvements in accuracy compared with traditional methods. OverFeat simplely applying CNN as a sliding window function on an image pyramid. and R-CNN introduced a region proposal-based strategy, where only the interested regions can be kept and scale-normalized before classifying with a CNN, which greatly improved compute efficiency. the recent proposed Fast R-CNN[7] and Faster R-CNN[8] methods achieved more accurate result and compute efficiency, In Faster R-CNN, the selective search of interested region is altered to a region proposal network, in which a objectness score map was generated by adding an extra convolutional layer to the base network and the region proposal box was generated by this score map. In this research we adopt this score map to help us locate the suspicious stuff in image.

2.3 View-based 3D shape recognition

There has been existing work on recognizing 3D shape with CNNs, some methods resort to CNNs to extract features from multiple 2D representations[3, 4], for there are a number of advantages of using 2D representation. One main reason is the trade-off between resolution and compute efficiency[3], for example, a 16×16×16 grid of 3D voxels input shares the same input size with a more refined 64×64 2D image input, so the multi-view CNN can capture more details. In addition, there are many other ideas. Charles R.Qi et al proposed the PointNet to directly process the original point cloud data and classify the model data using a RNN-like network[5]. The slice-based CNN method proposed by F.Gomez-Donoso takes slices of the 3D models as the CNN input which alleviates the resolution trade-off[14]. In this research, we adopt the multi-view based 3D shape descriptor and process each view of the models individually.

3 The Proposed Approach

In this section, we present the approach of this paper. In order to predict which part of body contains suspicious stuff, we propose a two stage algorithm. In the

first stage, a pre-trained deep CNN is used to extract score maps from images by different views of 3D human body, predicting objectness score maps at each view. Then in the second stage, we aggregate those maps into a single feature and use a regularized MLP network for detection. The complete architecture is depicted in Fig.1.

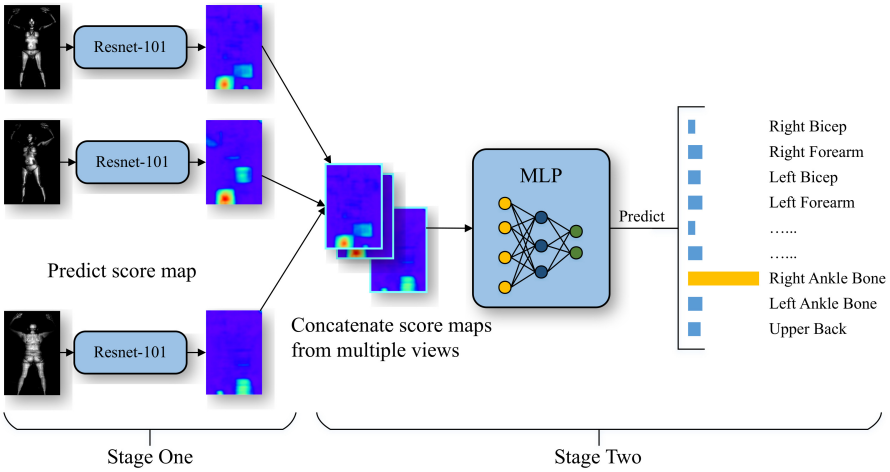


Fig. 1: The proposed two stage approach for object detection in 3D images.

3.1 Predict score map

We employ the VGG16[12] network and Resnet-101[11] network as the base network to predict a score map from each image input, which is in lower resolution than the input image. As in the original Faster R-CNN[8], the top fully connected layer in base network are removed, then RPN network is concatenated behind the lase convolution layer, in which the score map prediction and bounding box regression are performed. In this task we are not interested in the exact position of the suspicious stuff, so we remove the bounding box regression branch in RPN network.

For training RPNs, Faster R-CNN adopted anchor-based method in which every spatial location of score map was assigned an anchor box and classifiy each anchor box to positive or negative based on their overlap with ground-truth box. But actually in experiments we find this method is inclined to produce false alarm. In this scene we utilize CNN as probability regression function with the regression target of each position is defined as:

$$p(p|G) = \max_{g_i \in G} \begin{cases} 1 - (\frac{x_p - cx_i}{w_i})^2 (\frac{y_p - cy_i}{h_i})^2, & \text{if } p \text{ in } g_i \\ 0, & otherwise \end{cases} \tag{1}$$

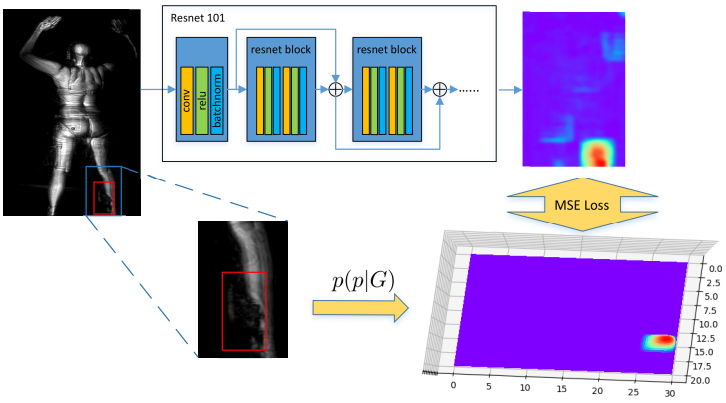


Fig. 2: In the first stage, we employ a single network to predict the score map of each views of images. we append a 1×1 convolution layer to base network, and a sigmoid activation layer to clamp the output value within $(0, 1)$.

Here, G is the set of ground-truth boxes and g_i is the i-th ground-truth box. x_p and y_p is the pixel position in score map w.r.t the input image. In the train phase, we apply the Mean-Square-Error(MSE) loss and Adam optimizer with learning rate of 10^{-4} at initial. The scheme of this stage is depicted in Fig.2.

3.2 Detect with Multi-view Representations

In the first stage, the score maps contain the position information of suspicious stuff in multi-view representations, we claim that those information is rich enough for us to locate the suspicious stuff in human body. In the second stage, we take score maps generated from multi-view representation of a single sample as input, and introduce a MLP network with softmax activation to produce the probabilities of each body zones. In experiments, we introduce lasso regulations to reduce over-fitting. So the target loss function is defined as:

$$L(x, l) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{z=1}^{M} \left[l_i^z \log f_z(x_i, \omega) + (1 - l_i^z) \log(1 - f_z(x_i, \omega)) \right] + \lambda \|\omega\|^2 \quad (2)$$

Where l_i^z denotes the label of the z-th zone in i-th sample, f is the MLP function while ω denotes the weight parameters, and λ is the weight of optional lasso penalty.

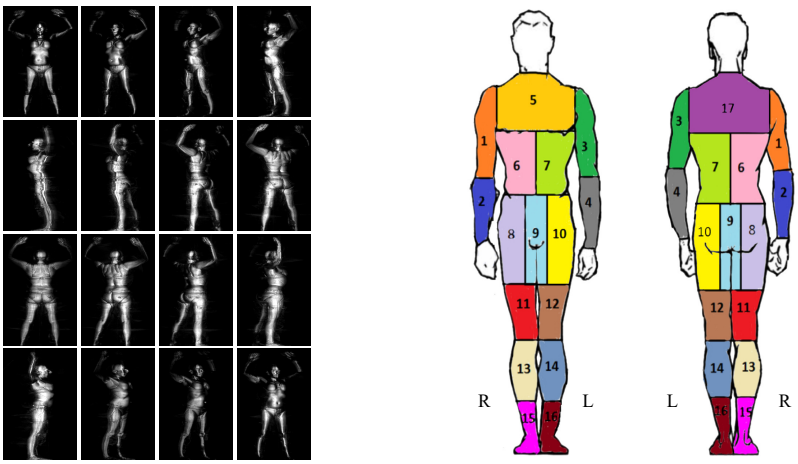
4 Experiments

In this section we provide the training details of our models, results achieved on the split validation set and test set result in the leaderboard.

4.1 Dataset

The Passenger Screening dataset used in this research comes from Kaggle competition at `https://www.kaggle.com/c/passenger-screening-algorithm-challenge`, This dataset contains thousands of body scans acquired by the HD-AIT system. The competition task is to predict the probability that a given body zone (out of 17 total body zones) has suspicious stuff present. For how much views is needed for efficient detection, we select 16 views as the same as in MV-CNN[3] which can balance between detection accuracy and compute efficiency. So we rendered each human body models in 16 different views which are equally spaced around the model in advance. Fig.3a shows the rendered images of a human body and Fig.3b shows the partition of body zones.

Since we only knew which part of body hiding suspicious stuff, the exact position of them is difficult to calculate based on the labels in the dataset. The stuff maybe occluded by human body in some views. In order to guide the algorithm to predict probabilities based on the suspicious stuff it find, and prevent the algorithm from over-fitting to the background, we hand-labeled the position of suspicious stuff in the train image set, which takes 80% of the total dataset and the validate set takes the rest 20%. The position of each suspicious stuff is annotated as a box with four scalars g_i=(cx_i,cy_i,w_i,h_i) which mark the center and size of the box.



(a) (b)

Fig. 3: Overview of passenger screening dataset. (a)A 3D sample from passenger screening dataset, which rendered in different views in 16 directions. (b)17 different zones of human body.

4.2 Training and Implementation

We training the two networks of different stages sequentially. In stage one, we construct VGG16 and Resnet-101 network and initialize them with ImageNet[2] pretrained weights. In order to make the model more robust to various object size, we apply random scaling, cropping and flipping to the input image, and resize the images to the fix size of 400×224 in both training and testing. After training the network, we fixed the network parameters and stacked a MLP classification network on it as described before. Then we optimized the target function in equation 2 until it converged.

We use Tensorflow as the deep learning framework, The models are trained on a single NVIDIA GTX1070 GPU within 12 hours.

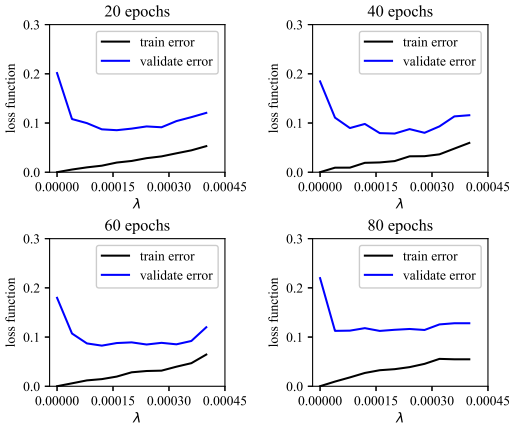


Fig. 4: the training and evaluating loss in different time of training progress and different penalty weight λ.

4.3 Results

Validation Data We validate our method using randomly split dataset for validation, which takes 20% of the total dataset. While training stage one, we save the model parameters after training 20, 40, 60 and 80 epochs, and we train stage two with those four parameters and different regulation parameter λ. The train error and validate error with respect to λ is depict in Fig.4, which indicates the network of stage one starts to overfit after training 40 epochs. It also shows that the MLP detection stage needs moderate regulation to reduce overfitting. In the end, we choose network parameters in 40-th epoch and $\lambda = 0.0001$ as the final model. The result is shown in Table.1, where we compare the result of using VGG16 or Resnet-101 as base network in stage one and the result of using

weight regulation or not in stage two. In all experiments we use 0.5 as positive threshold for calculating accuracy.

From Table.1, we find the model with Resnet-101 network and regulation achieves the best performance, where the model with VGG16 network is slightly worse. From this table, we can conclude that our approach can efficiently detect and locate the suspicious stuff in 3D human body.

Table 1: Results on validation data set

Method	Validate Loss	Accuracy
VGG16	0.220093	96.7263%
Resnet-101	0.179851	97.0844%
VGG16 + regulation	0.110845	97.4005%
Resnet-101 + regulation	**0.086878**	**97.9028**%

Test Data In order to test our model in practical application scenario, we applied our model to the test dataset which contains 2,000 new samples collected from the real world and uploaded the prediction result to kaggle server, The result from kaggle website is listed in Table 2. the test loss is sightly larger than validate loss, which demonstrated the effectiveness of our model.

Table 2: Results on test data set

Method	Validate Loss
VGG16 + regulation	0.166223
Resnet-101 + regulation	0.145261

5 Conclusion

We propose a two stage approach based on multi-view convolution neural network to detect and locate object in 3D images, and we apply our approach to detect suspicious stuff in 3D human scan images from airport security scanner. The experiments demonstrated the the proposed approach can efficiently detect and locate suspicious stuff in human body scans. Further work includes generalize this approach to different 3D images and various targets.

6 Acknoledgment

This research work is funded by the National Key Research and Development Project of China (2016YFB0801003) and the Sichuan province & university cooperation (Key Program) of science & technology department of Sichuan Province (2018JZ0050).

References

1. B. Elias, "Airport body scanners: The role of advanced imaging technology in airline passenger screening". *Congressional Research Service, Library of Congress*, 2012.
2. A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet classification with deep convolutional neural networks." in *Proc. Adv. Neurla Inf. Process. Syst. (NIPS)*, 2012, pp. 1097-1105.
3. H. Su, et al. "Multi-view Convolutional Neural Networks for 3D Shape Recognition." in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2016, pp. 945-953.
4. Qi, Charles R., et al. "Volumetric and multi-view cnns for object classification on 3d data." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2016, pp. 5648-5656.
5. Qi, Charles R., et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2017, pp 77-85.
6. R. Girshick, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2014, pp. 580-587.
7. R. Girshick. "Fast R-CNN." in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1440-1448.
8. S. Ren, K. He, R. Girshick, and J.Sun, "Faster R-CNN: Towards teal-Time object detection with region proposal networks." in *Proc. Adv. Neurla Inf. Process. Syst. (NIPS)*, 2015. pp. 91-99.
9. T. Lin, et al. "Feature pyramid networks for object detection." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2017, pp. 2117-2125.
10. J. Dai, et al. "R-fcn: Object detection via region-based fully convolutional networks." in *Proc. Adv. Neurla Inf. Process. Syst. (NIPS)*, 2016, pp. 379-387.
11. K. He, et al. "Deep residual learning for image recognition." in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2016, pp. 770-778.
12. K. Simonyan, and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015. [Online]. Available: https://arxiv.org/pdf/1409.1556
13. P. Sermanet, et al. "Overfeat: Integrated recognition, localization and detection using convolutional networks." in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014. [Online]. Available: https://arxiv.org/pdf/1312.6229
14. F. Gomez-Donoso, et al. "Lonchanet: A sliced-based cnn architecture for real-time 3d object recognition." *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 412-418.
15. O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional networks for biomedical image segmentation." in *Proc. Int. Conf. Med. Image Compput. Comput. Assist. Intervent. (MICCAI)*, 2015, pp. 234-241.