Progressive Image Recognition Method and its Application in Security Inspection Machines

Wu Jianxing¹, Zeng Dexin¹, Ju Qiaodan¹, Chang Zixuan¹ and Yu Hai¹,*

¹Software College, Northeastern University, Shenyang, Liaoning, China

Abstract: Background: Owing to the ability of a deep learning algorithm to identify objects and the related detection technology of security inspection equipment, in this paper, we propose a progressive object recognition method that considers local information of objects.

Methods: First, we construct an X-Base model by cascading multiple convolutions and pooling layers to obtain the feature mapping image. Moreover, we provide a “segmented convolution, unified recognition” strategy to detect the size of the objects.

Results: Experimental results show that this method can effectively identify the specifications of bags passing through the security inspection equipment. Compared with the traditional VGG and progressive VGG recognition methods, the proposed method achieves advantages in terms of efficiency and concurrency.

Conclusion: This study provides a method to gradually recognize objects and can potentially assist the operators in identifying prohibited objects.

Keywords: Progressive image recognition, convolutional neural network, security inspection, algorithm, X-ray, deep learning.

1. INTRODUCTION

Security inspection machines are widely used in airports, railway stations, and major exhibition halls. The primary role of a security inspection machine is to provide an image of objects inside a package so that the personnel can determine whether it contains dangerous items. However, slow manual inspection is inefficient and, hence, indirectly limits the throughput of the security entrance, resulting in congestions at the gate.

Modern security inspection machines mainly use X-ray technology to detect and identify dangerous goods. X-rays have some impressive properties, such as penetrability, fluorescent effects, and photographic effects. When X-rays pass through different materials, a black-and-white image can be formed on a screen according to the amount of X-ray absorbed by the object. These images provide detailed information about the scanned items, which can be used to identify whether these items are safe [1]. However, at present, the accuracy and efficiency of X-ray security screening depend entirely on the experience of the staff. Hence, it involves at least two problems.

The personnel may misjudge (i.e., identify a non-dangerous good as a non-dangerous one). Human errors are likely to occur in scenarios with high workload, poor working conditions, or complex items to be screened. It is reported [2] that the accuracy of personal inspection is appropriately between 80% to 90%.

A certain amount of judgment time is required. To reduce human errors, staff should be given enough time to judge, thereby reducing the efficiency of security screening.

In recent years, the use of deep learning algorithms to identify people and objects in images has received widespread attention from researchers. The backpropagation algorithm [3], proposed by Rumelhart and McClelland in 1986, and recent deep neural networks [4], proposed by Hinton and Salakhutdinov in 2006, laid a solid foundation for artificial intelligence theory. In particular, Convolutional Neural Networks (CNNs) [5-7] achieved excellent results in image recognition. Besides, detecting and recognizing objects in X-ray images [8-10] have been evolved with the development of CNNs. However, most existing CNNs cannot be directly applied to security screening.

First, during security screening, objects are scanned continuously, and images of different sizes are generated. However, existing deep CNNs required a fixed-size input image (e.g., 224×224) until He et al. [11] equipped networks with another pooling strategy, “spatial pyramid pooling,” in 2014. In 2016, they presented a residual learning framework [12] to ease the training of networks that were substantially deep-
er than those used previously. In a study [13], the authors developed an effective solution to the resulting nonlinear optimization problem without the need for stochastic gradient descent and accelerated the test-time computation of CNNs.

Another problem with continuous security screening is that images emerge progressively on the screen. Thus, we cannot extract the features of the global image immediately. A novel method to improve the flexibility of a CNN, which is called multi-scale orderless pooling (MOP-CNN), was proposed [14] to address this problem. MOP-CNN extracts partial image features and combines them to obtain global image recognition. It provides a roadmap of using partial image information to obtain image recognition.

Furthermore, current applications of CNNs are limited to known, static, and one-time inputs. However, people often recognize and understand objects in a gradual process. In this paper, as a case study, we address the problem of determining luggage size in a security inspection machine. Identifying the size of scanned objects beforehand can potentially enhance the identification efficiency of dangerous goods. Hence, fast identification of the size of the scanned object is essential for the rapid identification of dangerous goods.

In this study, we aim to design a new CNN architecture and operation process for progressive image recognition. Specifically, we propose the following solutions:

For manual identification, machine recognition is used to provide relevant prompts for personnel. The model can identify specifications of bags and/or luggage and remind the personnel which dangerous goods are likely to be contained in them. Thus, judgment workload is reduced, and human error rates can be decreased to a certain extent.

This method is used as a part of a semi-automatic security system, that is, an automatic security inspection is used together with manual judgment. In particular, we design a relevant image recognition algorithm that can automatically identify the presence of dangerous goods based on the scanning pattern of bags and manually determine whether the result is accurate. Subsequently, the workload is significantly reduced, thereby reducing the burden on the personnel. In this scheme, the system aims to reduce the target range of image recognition according to the specification and to improve the efficiency of detection.

The remainder of this paper is structured as follows. Section 2 describes the proposed method. Section 3 presents the experiments and results. Finally, Section 4 concludes the paper.

2. PROGRESSIVE IMAGE RECOGNITION METHOD

This section analyzes the local characteristics of the security checkbox and bag and presents a progressive object recognition method based on a CNN, which can rapidly analyze the size and scale of the box.

2.1. Base Model of X-ray Image Recognition

Compared with the GoogLeNet model [15], the VGG-16 network model [16] has a high recognition accuracy and strong transfer learning ability. In this section, we propose a simple and fast basic model for X-ray image recognition based on the VGG-16 model. We call this basic model as X-Base model.

The X-Base model uses a 3 × 3 convolution kernel in the VGG model and fixes the sliding step (stride) to 1 pixel. The space filling of the convolution layer (padding) is used to maintain the spatial resolution of the convoluted image. The kernels in MaxPool layers are 2 × 2 sliding windows with a sliding step of 2 pixels. All layers, except for the last fully connected classification layer, use the rectified linear unit (ReLU) as an activation function. In a research work [16], it has been proven that a stack with two 3 × 3 convolutional layers (without a pooling layer in the middle) is equivalent to an effective receptive field of 5 × 5. Moreover, three such layers are equivalent to an effective receptive field of 7 × 7. Despite having the same convolutional effective receptive field, using multiple convolutional layers allows the image data to undergo more nonlinear corrections through the activation function, such as ReLU. Therefore, the model can learn more complicated features, and its decision function is more discriminative. The X-Base model, proposed in this paper, inherits this advantage. Because there are only simple line images in the X-ray scanning image, the 1 × 1 convolution filtering operation in the VGG-16 model does not exist in our X-Base model. In addition, the proposed model significantly reduces the number of network layers to avoid overfitting owing to the characteristics of the X-ray scans, including small size and few features. Fig. (1) shows the X-Base model of the X-ray image recognition proposed in this section. Finally, we can obtain a map image with 128 channels.

Let $Z_i^m$ represent the $j$-th map image of the $m$-th layer network output. The number of input channels is denoted by $N$. Based on the construction of the X-Base model described above, we have

$$Z_j^m = f \left( \sum_{i=1}^{N} z_i^{m-1} \ast k_{ij}^m + b_j^m \right),$$

where $f(x)$ is the activation function, $K_{ij}^m$ is a convolution kernel parameter that maps the $i$-th map image to the $j$-th map image in the $m$-th layer neural network, $\ast$ represents the convolution operation, and $b$ is the offset.

2.2. Design of a Progressive Image Recognition Method

As the conveyor belt of a security inspection machine transmits bags into the X-ray range, X-ray images of the bags gradually appear on the screen. To simulate this process, we equally divide the X-ray image into $N$ parts ($N \geq 1$), and the different parts are gradually fed into the network model for training. According to the different stages of merging intermediate results, we propose the following progressive image recognition method.

(1) The input images collected in different stages are called "0%-10% image," "10%-20% image," ... , "90%-100% image." To enhance efficiency, our method reuses the X-Base model and the trained parameters to the extent possible. Thus, each stage image is input to the X-Base model to obtain the corresponding feature mapping image. Subsequently, the feature mapping image matrices are orderly merged into a larger matrix, as shown in Fig. (2).
Fig. (1). X-Base model. (A higher resolution / colour version of this figure is available in the electronic copy of the article).

Fig. (2). “Merge” operation diagram.

(2) The merged feature map image is converted into a vector through a flattening layer. Subsequently, the final prediction result can be further obtained through the processing of the fully connected layer. The detailed progressive recognition algorithm is described in Algorithm 1.

Algorithm 1. Progressive image recognition algorithm

**Input:** Image fragments \( I = (I_1, I_2, \ldots, I_{10}) \) obtained from the security inspection machine.

**Output:** \( R \in [0,1,2] \) represents the result of specification classification: small, medium, or large.

1: for every \( I \in I \) do
2: Let \( I \) pass through the X-Base model to obtain its corresponding feature mapping image segment \( Z_i \).
3: Stitch \( Z_i \) after the existing map image matrix and get the new map image matrix \( Z_i = [Z_{i-1}, Z(i)] \).
4: Take \( Z_i \) as the input of the fully connected network to obtain the result: \( R_i \).
5: end for

Fig. (3) illustrates the process of our proposed progressive recognition method. Apparently, this process is dynamic, and we can convolute and pool each image simultaneously. Therefore, compared with traditional image recognition models, our proposed method has not only higher parallelism but also better recognition efficiency. Moreover, it allows the
incremental addition of previous results and, hence, is highly scalable.

3. EXPERIMENTS

To verify the validity of the proposed method, we compared it with traditional complete image recognition methods. This section presents data acquisition, image preprocessing, and training.

3.1. Dataset Collection and Preparation

3.1.1. Data Acquisition

Although SIXray dataset [17] and GDXray [18] provide lots of images for baggage inspection, the aim is to detect dangerous objects. The baggage in SIXray and GDXray is not integrated. Owing to the limited number of images acquired by security inspection machines, we constructed a dataset by collecting images from the Internet and converting them into X-ray images. We crawled 201,735 luggage pictures from the electronic business platform called JD. According to the labels provided by the e-commerce platform, there are six main types of luggage: suitcases, travel bags, backpacks, shoulder bags, waist bags, and handbags. Considering the actual requirements of various bags, the relative specifications of the bags could be modified to a certain range. For example, the suitcase is no smaller than the backpack. Otherwise, it will lose practical value. Hence, we divided the existing six types of luggage into three categories, namely, large bags, medium bags, and small bags, according to specifications. Generally, large bags include suitcases and travel bags, medium bags include backpacks and shoulder bags, and small bags include waist bags and handbags. Thus, we created three folders (i.e., “Big Bag,” “Medium Bag,” and “Small Bag”) and loaded the six types of luggage image folders to the corresponding folders. The image file path was used to distinguish the specifications of various luggage images, which are the basis for making image tags.

3.1.2. Filtering Luggage Images

The background of the scanned images from the security inspection machine is always monotonous. Thus, complicated backgrounds in the crawled pictures (such as trademarks, models, and other irrelevant elements) need to be simplified. Therefore, in this section, we filter the 201735 luggage pictures obtained. We first used OpenCV [19] to read the luggage images and transform them into three-channel matrices. Subsequently, the “find contours” function provided by OpenCV was used to find the luggage locations. To simplify the process, we set a threshold $\delta$. For the parts outside the area of the bag, any image with the difference between two pixels exceeding the set threshold value $\delta$ was removed. Based on the above operations, we successfully filtered out solid color luggage images that matched the experimental background. According to the specification data provided on the JD e-commerce platform, the images whose bag specifications are significantly different from the category to which they belong are screened. We also set specific thresholds for the length, width, and height of the luggage of various specifications. All luggage images that do not match the thresholds are adjusted to the corresponding category for further filtering. Finally, manual inspection and screening were performed to ensure that a solid-color background and accurate classification of luggage images are obtained. In this step, 67,046 original images were obtained.

3.1.3. Image Preprocessing

There is an extreme difference between an object image captured by the X-ray scanner of the security inspection machine and an object image downloaded from the e-commerce platform. Therefore, the existing images must be preprocessed to simulate X-ray images. Through observation, we can find that the main content of the image scanned by the security inspection machine is the outline of the luggage composed of lines, most of which are filled with a black background. Therefore, we transformed the obtained images
to contour images with the same black background. The following preprocessing steps were used:

1. Reset image size.
2. Do “open” operation (i.e., corrosion before expansion) to blur them.
3. Transform the three-channel color object image into a single-channel grayscale image.
4. Recognize edges by a canny edge detector.
5. Divide every image evenly into ten parts.

As shown in Fig. (4), a three-channel color article image is converted into a 10-segmented single-channel line image, which is similar to the image obtained by an X-ray scanner in the security inspection machine.

3.2. Model Training and Testing

3.2.1. Dividing the Dataset

In this step, after shuffling 67,046 images in the original dataset, we first extracted 7046 copies as the verification set. Next, to ensure balanced training and testing of the model, we divided the remaining 60,000 copies into a training set and a test set according to a ratio of 9:1. The validation set was used to evaluate the model training, and the test set was used to evaluate the usability and generalization ability of the model. Because it can be guaranteed that the samples in the validation set are not used for training, the validation results have high reliability.

3.2.2. Model Training

In this study, 54,000 image samples were used for training. Because the images scanned by the security inspection machine mostly contain lines, the number of image features is relatively small. In this case, overfitting is likely to occur during the training. Therefore, we set the epoch parameter during the training to 2, the batch value to 32, the momentum to 0.9, and the fixed learning rate to 0.01.

During the training, we stored all samples of each segment (such as input 1) in its corresponding training matrix (such as I1) as a matrix. Moreover, we ensured that the subscripts of different segments of the same sample are consistent in their respective training matrices. In this way, 10 training matrices corresponding to the input sequence were obtained. The size of each matrix was 54,000 × 10% Z × Z. After constructing the model and configuring the parameters, we placed the training matrix of the first segment into the X-Base model to obtain the first segment map image Z1. Similarly, we obtained map images Z2, Z3, Z4, and Z5 of the 2nd–5th segments. Next, we stitched the map images of each segment to obtain a relatively complete map image Zmap = [Z1, Z2, Z3, Z4, Z5]. The prediction result P was obtained by a 50% Z × Z fully connected network. After calculating the loss function according to the real result y, the convolution kernel weight and other parameters were adjusted by back-propagation. After two rounds of training, the loss function has gradually converged. Finally, the training was completed.

3.3. Comparative Analysis

This section analyzes the proposed method in three aspects: accuracy, recall, and F1 score. Table 1 shows the confusion matrix of the classification.

As presented in Table 1, TP is defined as a correct match; FP represents the items that should be judged as wrong but are regarded as correct; FN represents the items that should be judged as correct but are regarded as wrong; TN is a correct non-match. Furthermore, the three indicators used in this section are specifically defined as follows:

$$\text{accuracy rate} = \frac{TP + TN}{TP + TN + FP + FN}$$ \hspace{1cm} (2)

$$\text{recall rate} = \frac{TP}{TP + FN}$$ \hspace{1cm} (3)

$$F_1 = \frac{2TP}{2TP + FP + FN}$$ \hspace{1cm} (4)

3.4. Model Testing

First, we restricted the input to 50% of the original image, set the parameters of the training function according to the values listed above, and trained our model using the training set. We obtained the accuracy rates of the training, testing, and validation sets as 93.62%, 90.93%, and 90.35%,
Table 1. Confusion matrix of the classification.

<table>
<thead>
<tr>
<th>Actually</th>
<th>Prediction</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (TP)</td>
<td>Actual Positive (TP+FN)</td>
</tr>
<tr>
<td>0</td>
<td>False Positive (FP)</td>
<td>Actual Negative (FP+TN)</td>
</tr>
<tr>
<td>Sum</td>
<td>Predicted Positive (TP+FP)</td>
<td>Predicted Negative (FN+TN)</td>
</tr>
</tbody>
</table>

Fig. (5). Model test results. *(A higher resolution / colour version of this figure is available in the electronic copy of the article).*

Table 2. Comparison of our model and VGG model.

<table>
<thead>
<tr>
<th></th>
<th>Our Model</th>
<th>VGG Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image percentage</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>90.35%</td>
<td>92.16%</td>
</tr>
<tr>
<td>Effectiveness</td>
<td>0.0383</td>
<td>0.0063</td>
</tr>
</tbody>
</table>

respectively, and the recall rates were 92.87%, 90.18%, and 89.84%, respectively; the F1 scores were 93.24%, 90.55%, and 90.09%, respectively, as shown in Fig. (5).

As shown in Fig. (5), compared with the training and verification sets, the experimental results of the model on the test set exhibited no significant difference in accuracy, recall, and F1 score. Thus, our proposed model possessed a good generalization ability.

To further analyze the performance of our methods, we define effectiveness as

\[
\text{Effectiveness} = \frac{\text{accuracy rate}}{\text{training time}}
\]

(5)

Thus, a higher accuracy and lower training time indicate that the model is more reliable and effective.

We further take the complete image as the input for VGG recognition. Table 2 compares the accuracy and effectiveness of our proposed model with that of the VGG model.

The results show that the accuracy of object size recognition using the VGG model for full images is 92.16%, which is only 1.81% higher than the accuracy rate (90.35%) by using our method with 50% part of the image as input. This is because VGG works on a dataset with more complete and rich image features. Owing to the segmentation of the image in progressive recognition, the size of the feature map image obtained after the convolution and pooling operations is smaller than that of the feature map image of the complete image, and the links between the images are weakened to a certain extent. However, the time required to obtain and process a 50% image is almost 50% less than that for a complete image. By substituting the accuracy rate and the training time into (5), it can be found that the effectiveness of our proposed model and VGG model is 0.0383 and 0.0063, respectively (in %/s); the former is approximately five times higher than the latter. Therefore, the gradual method proposed in this paper can significantly improve efficiency (the timeline diagram is shown in Fig. 6).
In the above experiments, we have proven that our model possesses excellent efficiency and accuracy compared with the traditional static model, but we have not directly proved its advantages as a progressive method. Therefore, to further prove the superiority of our model as a progressive method, we applied the VGG model as a progressive identification method in a simple way to conduct a group of comparative experiments.

In these comparative experiments, we designed CNN models corresponding to the input size of each time period during the gradual entry of goods into the security inspection machine. In each time period, the image of the object that has entered the body of the security inspection machine is scanned and input into the corresponding model to obtain the results. Specifically, we first assumed that the size of the complete image is \( Z \times Z \). When 10% of the object enters the security machine, the size of the obtained image is 10% \( Z \times Z \). The first judgment result \([Y_1]\) can be obtained by placing it into the neural network corresponding to the input size. Next, when 20% of the object enters the security machine, we obtained an image size of 20% \( Z \times Z \), and we placed it into the neural network corresponding to the input size to obtain the second judgment result of the object; then, the existing result is \([Y_1, Y_2]\). Similarly, we can obtain the judgment result \([Y_1, Y_2, Y_3, Y_4, Y_5]\). The support vector machine (SVM) algorithm provides higher accuracy and recall. For comparison, we take the temporary results represented by different dimensional vectors as input and use the SVM model [20] to identify the size of items. The progressive VGG process is shown in Fig. (7).
It can be found that the progressive VGG model could avoid information loss caused by cutting in our proposed model. Meanwhile, with its progressive characteristics, the VGG model could run tasks of multiple periods simultaneously. When there is enough probability of accepting the judgment result of the item specification, the object can no longer be detected to reduce the workload. Therefore, the efficiency of the model is greatly improved compared with the traditional static VGG model. Compared with the proposed model, the progressive VGG model makes full use of the information contained in the image. However, the disadvantages of this method are also evident. First, the image sizes used in each period are different. Owing to the different sizes of input images, we need to train and load multiple CNN models in the security machine simultaneously. In addition, compared with the other two methods, this method requires more SVM models. The requirement of multiple models increases the overall workload of model development. The storage and loading of more image data and model parameters also increase the hardware requirements for the security inspection machine. Moreover, because each image input into the neural network must be convoluted and pooled, and most of the content in the image has been operated in the same way in many previous periods, a large amount of time is wasted. This is the reason why it has a lower efficiency than our model.

The experiment was also carried out under the condition that 50% of the objects entered the security inspection machine. Through this experiment, we finally obtained the accuracy of the progressive VGG model in the testing set to be 90.84%, and its effectiveness was 0.01998, which was approximately half of our proposed X-Base model.

**CONCLUSION**

In this paper, we proposed a new CNN model based on the characteristics of X-ray images and compared it with the VGG model to identify the size of the luggage on security inspection machines progressively. We used parts of the X-ray images as input and stitched the partial maps generated by the proposed X-Base model. Finally, flattening layers and fully connected operations were performed to obtain the size of the scanned luggage. Compared with the VGG and progressive VGG models, our proposed model achieves a higher recognition efficiency with only a slight loss of accuracy. The simulation results verify the feasibility of our proposed method for image recognition using partial image features. However, owing to the limitation of datasets, we did not verify our proposed method on practical scanned images from security inspection machines. In the future, we will investigate on detecting prohibited objects in complicated environments, e.g., occlusion and rotation problems in cluttered bags.

**CONSENT FOR PUBLICATION**

Not applicable.

**AVAILABILITY OF DATA AND MATERIALS**

Not applicable.

**FUNDING**

This research was supported by the National Natural Science Foundation of China (Grant Nos. 61603082, 61977014, 61902056), the Fundamental Research Funds for the Central Universities (Grant Nos. N2017016, N2017011).
CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

We are grateful to Prof. Zhiliang Zhu for his helpful comments on the early versions of the paper. We also thank editors and reviewers for reviewing our manuscript and for their constructive comments to improve it.

REFERENCES