# Automated Deep Learning for Threat Detection in Luggage from X-ray Images

Alessio Petrozziello and Ivan Jordanov

[1] *University of Portsmouth, Portsmouth, U.K*
*Alessio.petrozziello@port.ac.uk, Ivan.Jordanov@port.ac.uk*

**Abstract.** Luggage screening is a very important part of the airport security risk assessment and clearance process. Automating the threat objects detection from x-ray scans of passengers' luggage can speed-up and increase the efficiency of the whole security procedure. In this paper we investigate and compare several algorithms for detection of firearm parts in x-ray images of travellers' baggage. In particular, we focus on identifying steel barrel bores as threat objects, being the main part of the weapon needed for deflagration. For this purpose, we use a dataset of 22k double view x-ray scans, containing a mixture of benign and threat objects. In the pre-processing stage we apply standard filtering techniques to remove noisy and ambiguous images (i.e., smoothing, black and white thresholding, edge detection, etc.) and subsequently employ deep learning techniques (Convolutional Neural Networks and Stacked Autoencoders) for the classification task. For comparison purposes we also train and simulate shallow Neural Networks and Random Forests algorithms for the objects detection. Furthermore, we validate our findings on a second dataset of double view x-ray scans of courier parcels. We report and critically discuss the results of the comparison on both datasets, showing the advantages of our approach.

**Keywords:** Baggage screening, Deep Learning, Convolutional Neural Networks, Image filtering, Object Detection Algorithms, X-ray Images.

## 1 Introduction

Identifying and detecting dangerous objects and threats in baggage carried on board of aircrafts plays important role in ensuring and guaranteeing passengers' security and safety. The security checks relay mostly on X-ray imaging and human inspection, which is a time consuming, tedious process performed by human experts assessing whether threats are hidden or occluded by other objects in a closely packed bags. Furthermore, a variety of challenges makes this process tedious, among those: very few bags actually contain threat items; the bags can include a wide range of items, shapes and substances (e.g., metals, organic, etc.); the decision needs to be made in few seconds (especially in rush hours); and the objects can be rotated, thus presenting a difficult to recognize view. Due to the complex nature of the task, the literature suggests that human expert detection performance is only about 80-90%

accurate [1]. Automating the screening process through incorporating intelligent techniques for image processing and object detection can increase the efficiency, reduce the time, and improve the overall accuracy of dangerous objects recognition.

Research on threat detection in luggage security can be grouped based on three imaging modalities: single-view x-ray scans [2], multi-view x-ray scans [3] [4], and computed tomography (CT) [5]. Classification performance usually shows improvements with the number of utilised views, with detection performance ranging from 89% true positive rate (TPR) with 18% false positive rate (FPR) for single view imaging [2] to 97.2% TPR and 1.5% FPR in full CT imagery [5].

The general consensus in the baggage research community is that the classification of x-ray images is more challenging than the visible spectrum data, and that direct application of methods used frequently on natural images (such as SIFT, RIFT, HoG, etc.) does not always perform well when applied to x-ray scans [6]. However, identification performance can be improved by exploiting the characteristics of x-ray images by: augmenting multiple views; using a coloured material image or employing simple (gradient) density histogram descriptors [7] [8] [9]. Also, the authors of [10] discuss some of the potential difficulties when learning features using deep learning techniques, on varying size images with out-of-plane rotations.

This work aims to develop a framework to automatically detect firearms from x-ray scans using deep learning techniques. The classification task focusses on the detection of steel barrel bores to determine the likelihood of firearms being present within an x-ray image, using a variety of classification approaches. Two datasets of dual view x-ray scans are used to assess the performance of the classifiers: the first dataset contains images of hand-held travel luggage, while the second dataset comprises scans of courier parcels. We handle the varying image size problem by combining the two views in one unique sample, while we do not explicitly tackle the out-of-plane rotation problem, instead, we rely on data augmentation techniques and on a dataset containing the threat objects recorded in different poses.

We investigate two deep learning techniques, namely Convolutional Neural Networks (CNN) and Stacked Autoencoders, and two widely used classification models (Feedforward Neural Networks and Random Forests) and the results from their implementation are critically compared and discussed.

**Fig. 1.** A sample image containing a steel barrel bores (top left cylinder in the top row) from the baggage dataset. The left image (in both rows) is the raw dual view x-ray scan, in the middle, the grey scale smoothed one, and on the right, the b/w thresholded one.

The remainder of the paper is organized as follows. Section 2 describes the datasets used in the empirical experimentation and illustrates the proposed framework; Section 3 reports details on the carried experiments and results; while conclusion and future work are given in Section 4.

## 2 Threat Identification Framework

The proposed framework for automated weapon detection consists of three modules: pre-processing, data augmentation and threat detection. The pre-processing stage comprises four steps: green layer extraction, greyscale smoothing, black and white (b/w) thresholding and data augmentation.

The original dataset consists of over 22000 images of which approximately 6000 contain a threat item (i.e., a whole firearm or a component). The threat images are produced by a dual view x-ray machine: one view from above, and one from the side. Each image contains metadata about the image class (i.e., benign or threat), and firearm component (i.e., barrel only, full weapon, set of weapons, etc). From the provided image library, a sample of 3546 threat images were selected containing a firearm barrel (amongst the other items), and 1872 benign images only containing allowed objects. The aim of the classification is to discriminate only the threat items - as common objects are displayed in both 'benign' and 'threat' samples (e.g., Figure 1 and Figure 2). During the pre-processing phase, each image is treated separately and the two views are combined before the classification stage.

The raw x-ray scans are imported in the framework as a 3-channel images (RGB) and scaled to 128x128 pixels in order to have images of same size for the machine learning procedure, and to meet memory constraints during training.

From the scaled image, the green colour channel is extracted as the one found to have the greatest contrast in dense material.

The resulting greyscale image is intended to reflect more accurately the raw x-ray data (i.e., measure of absorption). This step is performed to enable subsequent filtering and better identification of a threshold for dense material and eventually to facilitate the recognition of the barrel.

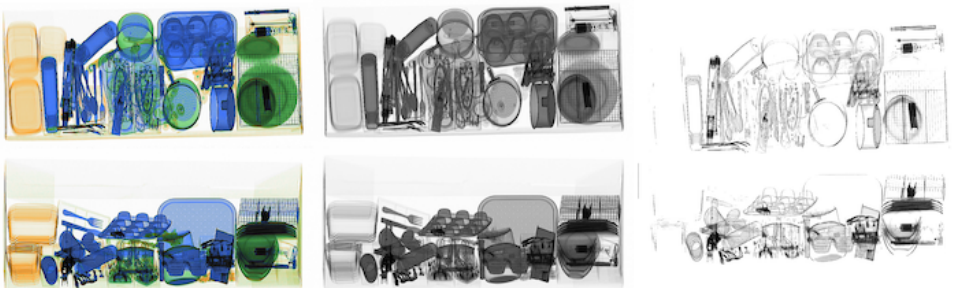

**Fig. 2.** A sample image containing a steel barrel bores (top right cylinder in the top row) from the parcel dataset. The left image (both rows) is the raw dual view x-ray scan, in the middle, the grey scale smoothed one, and on the right, b/w thresholded one. The parcel dataset usually contains a higher amount of steel objects and the barrels are better concealed.

A smoothing algorithm is applied on the greyscale image in order to reduce the low-level noise within it, while preserving distinct object edges. A number of smoothing algorithms were tested and a simple 3x3 kernel Gaussian blur was found to generate the best results. Then, on the smoothed image we apply a thresholding technique to isolate any dense material (e.g., steel). The chosen threshold is approximated within the algorithm to the equivalent of 2mm of steel, which ensures that metal objects, such as firearm barrels and other components ar e kept. This step removes much of the benign background information within the image, such as organic materials and plastics. The resulting image is normalised to produce a picture where the densest material is black and the image areas with intensity below the threshold are white. At this point, the instances for which the produced image lacks any significant dense material, can be directly classified as benign. From cursory examination of the operational benign test set, this is a significant proportion of the samples for the *baggage* dataset, while only filtering out a small portion of images on the *parcels* one (mainly because in the courier parcels there is a higher variety of big and small metallic objects compared to the hand-held travel luggage). When applying deep learning techniques on images, it is often useful to increase the robustness of the classification by adding realistic noise and variation to the training data (i.e., augmentation), especially in the case of high imbalance between the classes [11]. There are several ways in which this can be achieved: object volume scaling: scaling the object volume V by a factor v; object flips/shifts: objects can be flipped/shifted in the x or y direction to increase appearance variation. This way, for every image in the training set, multiple instances are generated, combining different augmentation procedures and these are subsequently used by the models during the learning phase. Lastly, the two views of each sample are vertically stacked to compose one final image (Figure 1 and Figure 2).

The four machine learning methods incorporated and critically compared in this work include two from the deep learning area, namely Convolutional Neural Networks (CNN) and Stacked Autoencoders; and two shallow techniques: Neural Networks and Random Forests.

The CNN are considered state-of-the-art neural network architectures for image recognition, having the best results in different applications, e.g.: from a variety of problems related to image recognition and object detection [12], to control of unmanned helicopters [13], x-ray cargo inspection [7], and many others. A CNN is composed of an input layer (i.e., the pixels matrix), an output layer (i.e., the class label) and multiple hidden layers. Each hidden layer usually includes convolution, activation, and pooling functions, and the last few layers are fully connected, usually with a softmax output function. A convolutional layer learns a representation of the input applying a 2D sliding filters on the image and capturing the information of contingent patches of pixels. The pooling is then is used to reduce the input size, aggregating (e.g., usually using a max function) the information learned by the filters (e.g., a 3x3 pixels patch is passed in the learned filter and the 3x3 output is then pooled taking the maximum among the nine values). After a number of hidden layers (performing convolution, activation, and pooling), the final output is flattened into an array and passed to a classic fully connected layer to classify the image.

Stacked Autoencoders, also called auto-associative neural networks, are machine learning technique used to learn features at different level of abstraction in an unsu-

pervised fashion. The autoencoder is composed of two parts: an encoder, which maps the input to a reduced space; and a decoder which task is to reconstruct the initial input from the lower dimensional representation. The new learned representation of the raw features can be used as input to another autoencoder (hence the name stacked). Once each layer is independently trained to learn a hierarchical representation of the input space, the whole network is fine-tuned (by performing backpropagation) in a supervised fashion to discriminate among different classes. In this work we use sparse autoencoders, that rely on heavy regularization to learn a sparse representation of the input.

## 3 Experimentation and Results

After the pre-processing and filtering off the images not containing enough dense material, we ended with 1848 and 1764 samples for classification of the *baggage* and *parcel* datasets respectively. The *baggage* dataset comprises 672 images from the benign class and 1176 containing threats; while the *parcel* dataset 576 and 1188 samples for the benign and threat classes respectively. Each dataset was split in 70% for training and 30% as independent test set. Due to their different operational environments, the baggage and parcel scans were trained and tested separately.
In this experiment we used a three layer stacked autoencoder with 200, 100, 50 neurons respectively, followed by a *softmax* output function to predict the classes probability. For the CNN we emploed a topology with three convolutional layers (with 128, 64 and 32 neurons) followed by a fully connected neural network and a *softmax* output function.
The RF was trained with 200 trees while the shallow NN had a topology of ***n-n-2***, where ***n*** was the input size. Since both RF and shallow NN cannot be directly trained on raw pixels, a further step of feature extraction was performed. In particular, we used histograms of *oriented Basic Image Features* (*oBIFs*) as a texture descriptor (as suggested in [6]), which has been applied successfully in many machine vision tasks. The *Basic Image Features* is a scheme for classification of each pixel of an image into one of seven categories, depending on local symmetries. These categories are: flat (no strong symmetry), slopes (e.g., gradients), blobs (dark and bright), lines (dark and bright), and saddle-like. *Oriented BIFs* are an extension of the *BIFs,* that include the quantized orientation of rotationally asymmetric features [14], which encode a compact representation of images. The *oBIF* feature vector is then fed as input into the RF and the shallow NN classifiers.

To evaluate the classification performance we employ three metrics: area under the ROC curve (AUC), the false positive rate at 90% true positive rate (FPR@90%TPR), and the F1-score. The AUC is a popular metric for classification tasks and the FPR@90%TPR is one cut-off point from the AUC, which describes the amount of false positives we can expect when correctly identifying 90% of all threats. The cut-off at 90% is suggested by [6] for the classification of x-ray images in a similar context. The F1-score is also a widely used metric for classification of imbalanced datasets that takes into account the precision (the number of correctly identified threats

divided by the number of all threats identified by the classifier) and the recall (the number of correctly identified threats divided by the number of all threat samples).

**Table 1** Baggage dataset results for the AUC, FPR@90%TPR and F1-Score metrics. The results are reported for the four classification techniques and three pre-processing step: raw data, grey scale smoothing and b/w thresholding.

| Metric | Technique | Raw | Smoothing | B/w thresholding |
|---|---|---|---|---|
| AUC | CNN | **93** | **95** | **96** |
| | Autoencoder | 75 | 78 | 90 |
| | oBIFs + NN | 85 | 87 | 94 |
| | oBIFs + RF | 66 | 72 | 80 |
| FPR @ 90% TPR | CNN | **9** | **7** | **6** |
| | Autoencoder | 70 | 60 | 26 |
| | oBIFs + NN | 50 | 31 | 14 |
| | oBIFs + RF | 86 | 66 | 53 |
| F1-Score | CNN | **91** | **93** | **93** |
| | Autoencoder | 60 | 65 | 81 |
| | oBIFs + NN | 64 | 67 | 79 |
| | oBIFs + RF | 36 | 41 | 56 |

As it can be seen from Table 1, the CNN outperformed the other methods with AUC ranging between 93% and 96%, depending on the pre-processing stage. The second best method was the shallow NN with AUC values between 85% and 94%, while the worst performance was achieved by the RF with 66%-80% AUC. Similar results were achieved when considering the FPR@90%TPR and F1-score metrics. The CNN reached the best FPR (6%) when trained on the b/w thresholded images, while still having only 9% FPR when using raw data. On the other hand, while achieving 14% FPR with the last stage of pre-processing, the NN performance dropped drastically when employing the raw and the smoothed data, with 50% and 31% FPR respectively. The same can be observed when using the F1-score: the CNN achieving up to 93%, followed by the Stacked Autoencoders and the shallow NN with 81% and 79% respectively. Once again, it is worth noticing that the CNN was the only technique able to score high classification accuracy across all used pre-processing approaches, while the other methods needed more time spent on the features engineering and extracting steps.

**Table 2** Parcel dataset results for the AUC, FPR@90%TPR and F1-Score metrics. The results are reported for the four classification techniques and three pre-processing step: raw data, grey scale smoothing and b/w thresholding

| Metric | Technique | Raw | Smoothing | B/w Thresholding |
|---|---|---|---|---|
| AUC | CNN | **80** | **79** | **84** |
| | Autoencoder | 65 | 66 | 75 |
| | oBIFs + NN | 65 | 69 | 84 |
| | oBIFs + RF | 63 | 63 | 79 |
| FPR @ 90% TPR | CNN | **46** | **46** | **37** |
| | Autoencoder | 66 | 69 | 70 |
| | oBIFs + NN | 71 | 75 | 40 |
| | oBIFs + RF | 91 | 88 | 56 |
| F1-Score | CNN | **86** | **83** | **87** |
| | Autoencoder | 40 | 43 | 55 |
| | oBIFs + NN | 36 | 32 | 63 |
| | oBIFs + RF | 34 | 42 | 58 |

Table 2 shows the performance metrics on the *parcel* dataset, illustrating generally lower performance across all techniques. This can be explained by the larger variety of metal items contained in the courier parcels, when compared to the objects contained in a hand-held airport luggage. Again, the CNN outperformed the other considered methods, with an AUC ranging from 79% to 84%, followed by the NN with 65% to 84%, RF with 63% to 79%, and the Stacked Autoencoders with 65% to 75%. The AUC achieved on the *parcel* dataset by the shallow NN, RF and Stacked Autoencoders are much closer than those achieved on the *baggage* one, where the best performing method outstands more.

Yet again, the CNN achieved the lowest FPR (37%), followed by the shallow NN with 40% FPR, the RF with 56% FPR and the Stacked Autoencoders with 70% FPR. Lastly, the F1-score metric produced the largest difference in values across the methods, with the CNN achieving up to 87% F1-score, followed by shallow NN with 63%, RF with 58% and Stacked Autoencoders with 55%. Also, in this case the CNN was the only technique able to classify threats with high accuracy, just using the raw images, where all other techniques performed very poorly (e.g., the AUC on raw data for the CNN was 15 percentage points better than the NN, while holding similar performance on the b/w thresholded one; 20 percentage points better in FPR@90% TPR when compared to the second best (Autoencoder); and even 46 percentage points better than the Autoencoder for the F1-score).

## 4 Conclusion

In this work we investigated a deep learning framework for automated identification of steel barrel bores in datasets of X-ray images in operational settings such as airport security clearance process and courier parcel inspections. In particular we compare two deep learning methods (Convolutional Neural Networks and Stacked Autoencoders), and two widely used classification techniques (shallow Neural Networks and Random Forest) on two datasets of X-ray images (*baggage* and *parcel* datasets). We evaluated the methods performance using three commonly accepted metrics for classification tasks: area under the ROC curve (AUC), the false positive rate at 90% true positive rate (FPR@90%TPR), and the F1-score. The obtained results showed that the CNN is not only able to consistently outperform all other compared techniques over the three metrics and on both datasets, but it is also able to achieve good prediction accuracy when using the raw data (whether the other techniques need multiple steps of data pre-processing and feature extraction to improve their performance). Furthermore, the CNN also achieved higher accuracy than the reported in literature results from human screening [1] (although, the employed datasets have not been screened by human experts, so an accurate direct comparison cannot be reported). Future work will explore application of different architectures for the CNN and Stacked Autoencoders, based on simulations on larger datasets to further investigate the result of this initial experimentation.

## References


1. S. Michel, S. M. Koller, J. C. de Ruiter, R. Moerland, M. Hogervorst and A. Schwaninger, "Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners," in *41st Annual IEEE International Carnahan Conference on Security Technology*, 2007.
2. V. Riffo and D. Mery, "Active X-ray testing of complex objects," *Insight-Non-Destructive Testing and Condition Monitoring,* vol. 54, no. 1, pp. 28-35, 2012.
3. D. Mery, V. Riffo, U. Zscherpel, G. Mondragon, I. Lillo, I. Zuccar, H. Lobel and M. Carrasco, "The database of X-ray images for nondestructive testing," *Journal of Nondestructive Evaluation,* vol. 34, no. 4, pp. 1-12, 2015.
4. D. Mery, V. Riffo, I. Zuccar and C. Pieringer, "Automated X-ray object recognition using an efficient search algorithm in multiple views," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013.
5. G. Flitton, A. Mouton and T. Breckon, "Object classification in 3D baggage security computed tomography imagery using visual codebooks," *Pattern Recognition,* vol. 48, no. 8, pp. 2489-2499, 2015.
6. N. Jaccard, T. W. Rogers, E. J. Morton and L. D. Griffin, "Tackling the X-ray cargo inspection challenge using machine learning," in *Anomaly Detection and Imaging with X-Rays (ADIX)*, 2016.
7. T. W. Rogers, N. Jaccard and L. D. Griffin, "A deep learning framework for the automated inspection of complex dual-energy x-ray cargo imagery," in *Anomaly Detection and Imaging with X-Rays (ADIX) II*, 2017.

8. G. Li and Y. Yu, “Contrast-oriented deep neural networks for salient object detection,” *IEEE transactions on neural networks and learning systems,* vol. 29, no. 1, pp. 6038-6051, 2018.
9. Y. Shen, R. Ji, C. Wang, X. Li and X. Li, “Weakly supervised object detection via object-specific pixel gradient,” *IEEE transactions on neural networks and learning systems,* vol. 29, no. 1, pp. 5960-5970, 2018.
10. M. Bastan, W. Byeon and T. M. Breuel, “Object recognition in multi-view dual energy x-ray images,” in *BMVC*, 2013.
11. C. Zhang, K. C. Tan, H. Li and G. S. Hong, “A Cost-Sensitive Deep Belief Network for Imbalanced Classification,” *IEEE transactions on neural networks and learning systems,* vol. 30, no. 1, pp. 109-122, 2019.
12. Z.-Q. Zhao, P. Zheng, S.-T. Xu and X. Wu, “Object Detection With Deep Learning: A Review,” *IEEE Transactions on Neural Networks and Learning Systems,* pp. 1-21, 2019.
13. Y. Kang, S. Chen, X. Wang and Y. Cao, “Deep Convolutional Identifier for Dynamic Modeling and Adaptive Control of Unmanned Helicopter,” *IEEE transactions on neural networks and learning systems,* vol. 30, no. 2, pp. 524-538, 2019.
14. A. J. Newell and L. D. Griffin, “Natural image character recognition using oriented basic image features,” in *International Conference on Digital Image Computing Techniques and Applications*, 2011.