TRANSFER LEARNING USING CONVOLUTIONAL NEURAL NETWORKS FOR OBJECT CLASSIFICATION WITHIN X-RAY BAGGAGE SECURITY IMAGERY

Samet Akçay[1], *Mikolaj E. Kundegorski*[1], *Michael Devereux*[2], *Toby P. Breckon*[1]
[1]Durham University, Durham, UK [2]University of Bristol, Bristol, UK

ABSTRACT

We consider the use of transfer learning, via the use of deep Convolutional Neural Networks (CNN) for the image classification problem posed within the context of X-ray baggage security screening. The use of a deep multi-layer CNN approach, traditionally requires large amounts of training data, in order to facilitate construction of a complex complete end-to-end feature extraction, representation and classification process. Within the context of X-ray security screening, limited availability of training for particular items of interest can thus pose a problem. To overcome this issue, we employ a transfer learning paradigm such that a pre-trained CNN, primarily trained for generalized image classification tasks where sufficient training data exists, can be specifically optimized as a later secondary process that targets specific this application domain. For the classical handgun detection problem we achieve 98.92% detection accuracy outperforming prior work in the field and furthermore extend our evaluation to a multiple object classification task within this context.

Index Terms— Convolutional neural networks, transfer learning, image classification, baggage X - ray security

1. INTRODUCTION

X-ray baggage security screening is widely used to maintain aviation and transport security, itself posing a significant image-based screening task for human operators reviewing compact, cluttered and highly varying baggage contents within limited time-scales. Within both increased passenger throughput in the global travel network and an increasing focus on wider aspects of extended border security (e.g. freight, shipping postal), this posed both a challenging and timely automated image classification task.

Fig. 1: Exemplar X-ray baggage imagery containing firearms.

Prior work on object detection in x-ray baggage imagery is limited. Aviation security screening systems that are available commercially include X-ray, CT and computer aided detection (to aid human screeners) that performs enhancement, segmentation and classification of baggage objects [1]. Handgun detection is investigated in [2] by training fuzzy k-NN classifier with shape context descriptor [3] and Zernike moments [4], but with limited evaluation over only 15 image examples.

The work of [5] considers the concept of bag of visual words (BoW) within X-ray baggage imagery using Support Vector Machine (SVM) classification with several feature representations (DoG, DoG+SIFT, DoG+Harris) achieving performance of 0.7, 0.29, 0.57 recall, precision and average precision, respectively. Turcsany et al. [6] followed a similar approach and extended the work presented in [5]. Using a bag of visual words with SURF feature descriptors and SVM classifier together with a modified version of codebook generation yields 99.07% true positive, and 4.31% false positive rates. BoW approach with feature descriptor and SVM classification is also used in [7] for the classification of single and dual view X-ray images. Best average precisions achieved for guns and laptops are 94.6% and 98.2% [7]. Inspired by implicit shape models, Mery [8] proposes a method that automatically detects X-ray baggage objects. By using visual vocabulary, occurrence structures and 200 X-ray bag images 99% and 0.2% true positive and false positive rates are achieved for handgun detection.

Baştan thoroughly reviews the current literature in his latest work [9], on which he studies applicability and efficiency of sparse local features on object detection in baggage imagery. This work also investigates how material information given in X-ray imagery and multi-view X-ray imaging affect detection performance, concluding that possible future work may use convolutional neural networks.

Motivated by [6], and current trends in convolutional neural networks (CNN), we propose a method that accurately classifies baggage objects by type. Unlike [6], in which the classical bag of visual words (BoW) is used with Speeded-Up Robust Features (SURF) and Support Vector Machine (SVM) classification, we employ a CNN approach for the entire feature extraction, representation and classification process. More specifically, with the use of a transfer learning [10] approach, we optimize the CNN structure designed by Krizhevsky *et. al.* [11] by fine-tuning its convolutional and fully-connected layers for the full feature to classification pipeline within this problem domain. To the best of our knowledge, this is the first study introducing deep convolutional networks[11, 12] to the X-ray baggage screening

problem.

2. CLASSIFICATION

Automated threat screening task in X-ray baggage imagery can be considered as a classical image classification problem. Here we address this task using the approach of transfer learning and convolutional neural networks based on the prior work of [10, 13, 11, 12].

2.1. Convolutional Neural Networks

Deep convolutional neural networks can be considered modernized version of multi layer perceptrons. They have been widely used in diverse fields such as speech recognition [14] and natural language processing [15], also becoming state of the art within computer vision for challenging tasks such as image classification [11], object detection [16] and segmentation [17]. Recent developments and affordability of GPUs and accessibility of large data sets have provided researchers with further insight into larger and more complex (deeper) network models [11]. Unlike the traditional neural networks with conventionally one or two hidden layers, CNN can include many more hidden layers [18, 19, 12]. Designing a CNN with certain number of layers can be application, data or designer dependent. Modern CNN include the following layers with varying characteristics: convolutional layers (feature extraction), fully connected layer (intermediate representation), pooling layer (dimensionality reduction) and non linear operators (sigmoid, hyperbolic functions and rectified linear units).

A key differentiator is that CNN is based on two main concepts named local receptive fields and shared weights [20]. Local receptive fields are small regions inside the image which provide local information with region size defined as a variable parameter. Similar to the notion of sliding window, local receptive fields are spread across the image such that each forms a neuron in the following hidden layer. Using shared weights and biases for neurons in hidden layers of CNN is another unique notion that provides many advantages. First of all, since each neuron in a hidden layer uses same weight and bias, hidden layers have distinct feature characteristics. In [13], for instance, it has been shown that first convolutional layers behave like Gabor filters. Having many convolutional layers gives one a very broad feature matrix. Another advantage of using shared weights is that total number of parameter used rapidly decreases, which gives us not only faster training times but also the opportunity to construct more complex (deeper) network structures. Even though using shared weights significantly decreases the number of parameters present, these still considerably exceed those of more traditional machine learning approaches (requiring specialist training regimes: [11]).

This high-level of parametrization, and hence representational capacity, make CNN susceptible to over-fitting in the traditional sense. To overcome this issue, a number of techniques are employed to ensure generality of the learned parameterization of the target problem. Within the network, convolutional layers are usually followed by pooling layers which down-samples the current representation (image) and hence reduces the number of parameters carried forward inaddition to improving overall computational efficiency. Furthermore the use of dropout, whereby hidden neurons are randomly removed during the training process, is used to avoid over-fitting such that performance dependence on individual network elements is reduced in favor of collective error reduction and representational responsibility for the problem space. In addition, with the use of the generalized technique called transfer learning, initial CNN parameterization (training) towards a generalized object classification task can then be further optimized (fine tuned) towards a specific sub-problem with related domain characteristics.

2.2. Transfer Learning

Modern CNN approach typically include varying number of layers (3-22) within their structure, leading to a human-like measurable performance in image classification tasks [21]. Presently, such networks are designed manually with the resulting parametrization of the networks performing training using a stochastic gradient descent approach with varying parameters such as batch size, weight decay, momentum and learning rate over a huge data set (typically 10^6 in size). Current state of the art CNN models as such designed by Krizhevsky *et. al.* [11], Zeiler *et. al.* [22], Szegedy *et. al.* [12], Simonyan *et. al.* [19] are trained on a huge dataset such as ImageNet [23] which contains approximately a million of data samples and 1000 distinct class labels. However, the limited applicability of such training and parameter optimization techniques to problems where such large datasets are not available gives rise to the concept of transfer learning [16, 24]. The work of [13] illustrated that that each hidden layer in a CNN has distinct feature representation related characteristics among of which the lower layers provide general features extraction capabilities (akin to Gabor filters and alike), whilst higher layers carry information that is increasingly more specific to the original classification task. This finding facilitates the verbatim re-use of the generalized feature extraction and representation of the lower layers in a CNN, whilst higher layers are fine tuned towards secondary problem domains with related characteristics to the original. Using this paradigm, we can leverage the *a priori* CNN parametrization of an existing fully trained network, on a generic 1000+ object class problem [21], as a starting point for optimization towards to the specific problem domain of limited object class detection within X-ray images. Instead of designing a new CNN with random parameter initialization we instead adopt a pretrained CNN and fine tune its parameterization towards our specific classification domain. Specifically, we make use of the CNN configuration designed by Krizhevsky *et al.* [11], having 5 convolutional layers, 3 fully-connected layer with 60 million parameters, 650,000 neurons, and trained over the

ImageNet dataset on an image classification problem in the ILSVRC-2012 competition (denoted as AlexNet). We also employ the network structure proposed by Szegedy *et al.* [12], which won the ILSVRC 2014 competition (denoted as GoogLeNet). The network is designed using many more layers (22) with 12 times fewer network parameters compared to AlexNet. From this point we then perform fine-tuning approach to the networks to train over the X-ray baggage dataset using propagation algorithm with stochastic gradient descent method. To observe the effect of input dataset dissimilarity, we freeze the parameters of certain layers, meaning that the pre-trained parameters are used for learning the new dataset instead of being updated during training. Training and testing are performed via the use of Caffe [25], a deep learning tool designed and developed by the Berkley Vision and Learning Center.

2.3. Application to X-ray Security Imagery

To investigate the applicability of CNN transfer learning in object classification X-ray baggage imagery, we address two specific target problems:- a) a two class firearm detection problem (i.e. gun Vs. no gun) akin to that of the prior work of [6] and; b) a multiple class X-ray object classification problem (6 classes: firearm, firearm-components, knives, ceramic knives, camera and laptop). Our data-set (6997 X-ray images) are constructed using single conventional X-ray imagery with associated false color materials mapping (from dual energy, [26] see Figure 1 and 2). To generate a dataset for firearm detection, we manually crop baggage objects, and label each accordingly (e.g. Figure 2) - on the assumption an in-service detection solution would perform scanning window search through the whole baggage image. In addition to manual cropping, we also generate a set of negative images by randomly selecting image patches from a large corpus of baggage images that do not contain any target objects. Following these approaches, as shown in Figure 2, we create a dataset for firearm detection with 17,419 samples (3924 positive; 13,495 negative). For the multiple class problem we separate firearms and firearm sub-components into two distinct classes. Similarly, regular knives and ceramic knives are considered as two distinct objects. Following the same procedure we generate a dataset with 9123 samples (firearm: 2847, firearm components: 1060, knives: 2468 ceramic knives: 1416, camera: 432, laptop: 900).

Evaluation of our proposed approach is performed against the prior SVM-driven work of Turcsany *et. al.* [6] and the use of Random Forest classification [27] within a similar bag of visual words framework. SVM is trained using grid search and *k*-fold cross validation routine optimizing parameters cost C, where $log_2 C \in \{-5, .., 15\}$ and kernel γ, where $log_2 \gamma \in \{-15, .., 3\}$ for bag of visual words of vocabulary sizes 500, 1000, 1500 and 2000 with the use of LIBSVM [28]. The SVM classifier is trained using RBF Kernels with $C = 8$ and $\gamma = 8$. Similar to the approach followed within the SVM framework, parameter grid search is performed for the best parameters of random forest of up to 1000 trees, adjusting sample count $\{2, .., 15\}$ and depth $\{5, .., 30\}$ for BoW vocabularies of 500, 1000, 1500 and 2000 feature words. Optimal performance was achieved with a random forest configuration of 1000 trees with a maximal depth of 15 and maximal sample count of 18.

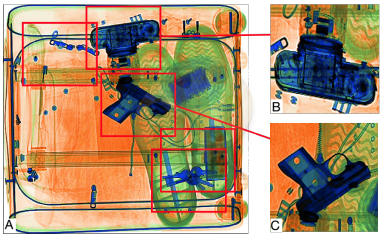

Fig. 2: Exemplar X-ray baggage image (A) with extracted data set regions for camera (B) and firearm (C) objects.

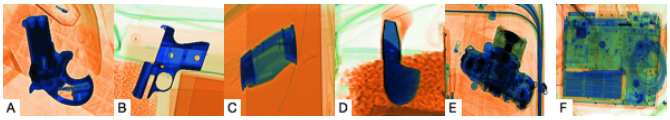

Fig. 3: Bag of visual words approach for multi-class problem. Type of baggage objects and the number of samples in our dataset is as follows: (A) Guns, (B) Gun Components, (C) Knives, (D) Ceramic Knives, (E) Cameras, (F) Laptops

3. EVALUATION

The performance of the proposed method and the prior work is evaluated by comparing the following metrics: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision (PRE), Recall (REC) and Accuracy (ACC).

Results for the two class problem is given in Tables 1 and 2, each of which are divided into two sections: - first section lists the performance of the CNN, notated as $AlexNet_{a_b}$, meaning that the network is fine-tuned from layer a to layer b, and rest of the layers are frozen. This means, for instance, $AlexNet_{4-8}$ is trained by fine-tuning the layers $\{4, 5, 6, 7, 8\}$ and freezing the layers $\{1, 2, 3\}$ (i.e. remain unchanged from [11]).

Table 1 shows the performance results of gun detection based on the training set. We see that true positives and true negatives have a general trend to decrease as the number of fine-tuned layers reduce. False positives and false negatives concordantly increases. Likewise, freezing more layers lowers the accuracy of the models. A conclusion can be reached from these results that fine tuning higher level layers and freezing lower ones have detrimental impact on the performance of the CNN models. This stems from the fact that features extracted from lower layers of the network are more

general, while the higher layers provide more specific information in regards to the training data. SVM has a competitive true positive rate of 97.43. However, suffering from high false positives of 14.93% results in poor performance compared to CNN. Similar to SVM, random forest performs well on precision and recall, yet high false positives rate cause worse accuracy compared to the rest of the models.

	TP%	TN%	FP%	FN%	PRE	REC	ACC
$AlexNet_{1-8}$	97.56	99.31	0.68	2.43	0.98	0.98	**0.99**
$AlexNet_{2-8}$	98.53	97.60	2.40	1.47	0.83	**0.99**	0.98
$AlexNet_{3-8}$	**98.62**	**99.79**	**0.21**	**1.38**	**0.99**	**0.99**	**0.99**
$AlexNet_{4-8}$	97.62	98.79	1.21	**1.38**	**0.99**	0.98	0.98
$AlexNet_{5-8}$	97.47	99.72	0.28	2.53	**0.99**	0.97	0.98
$AlexNet_{6-8}$	96.21	99.27	0.73	3.79	0.98	0.96	**0.99**
$AlexNet_{7-8}$	94.49	96.35	3.65	5.51	0.75	0.94	0.96
$AlexNet_{7-8}$	95.64	99.07	0.93	4.36	0.97	0.96	0.98
$AlexNet_{8}$	93.58	97.96	2.03	6.42	0.93	0.94	0.97
$SURF + RF$	94.10	65.44	34.56	5.90	0.90	0.94	0.87
$SURF + SVM$[6]	97.43	85.07	14.93	2.57	0.96	0.97	0.95

Table 1: Performance for the two class problem (Guns vs Non-Guns) using training set.

Table 2 shows the results of the models tested over the unseen dataset containing distinct type of objects that are never trained on the models. We see correlative performance to Table 1 such that as the number of fine-tuned layers decreases, performance of the CNN is adversely affected. SVM shows TP of 85.81% with a relatively high false positive rate of 11.76%. Even though SVM has the highest precision, its accuracy performs worse than any CNN. Furthermore, all of the CNN solutions consistently offer a lower FP and FN rate than the SVM or RF approaches (Table 1 - 2).

	TP%	TN%	FP%	FN%	PRE	REC	ACC
$AlexNet_{1-8}$	**99.26**	95.92	4.08	**0.74**	0.74	**0.99**	0.96
$AlexNet_{2-8}$	98.53	97.60	2.40	1.47	0.83	**0.99**	**0.98**
$AlexNet_{3-8}$	96.32	**97.81**	**2.19**	3.68	0.84	0.96	0.98
$AlexNet_{4-8}$	95.59	97.04	2.96	4.41	0.79	0.96	0.97
$AlexNet_{5-8}$	98.16	95.32	4.68	1.84	0.71	0.98	0.96
$AlexNet_{6-8}$	96.32	94.85	5.15	3.68	0.69	0.96	0.95
$AlexNet_{7-8}$	94.49	96.35	3.65	5.51	0.75	0.95	0.96
$AlexNet_{8}$	95.22	95.79	4.21	4.78	0.73	0.95	0.96
$SURF + RF$	80.74	67.28	32.72	19.26	0.95	0.81	0.79
$SURF + SVM$[6]	85.81	88.24	11.76	14.19	**0.98**	0.86	0.86

Table 2: Performance for the two class problem using test set.

Second set of experiments is based on the classification of multiple baggage objects, a more complex six class object problem. Here the lesser performing SVM and RF models are not considered (Table 1 - 2), in favor of the CNN approach. Instead, we only fine-tune two CNN structures by Krizhevsky *et. al.* (AlexNet) [11] and Szegedy *et. al.* (GoogLeNet) [12] to evaluate the feasibility of CNN for this problem domain. Performance is evaluated based on mean average precision (mAP) [29]. Figure 4 depicts per-class accuracy obtained via the use of GoogLeNet tested on randomly chosen dataset. Table 3 shows the overall performances of each of the model. Both show strong results for the multi class problem. AlexNet performs best when classifying laptops (99.70%). On the other hand, classifying gun components is a challenging task for AlexNet as it performs relatively worse (89.64%), stemming from the high visual overlap between classes. GoogLeNet shows strong performance even for the classes similar to each other (Gun / Gun Components, Knives / Ceramic Knives), and overall achieves superior mAP.

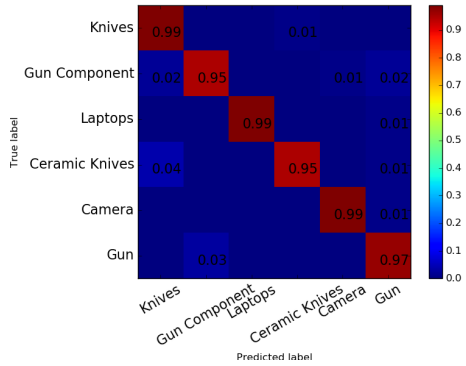

Fig. 4: Normalized confusion matrix of the fine-tuned GoogLeNet model tested on unseen test dataset.

	Camera	Laptop	Gun	Gun Component	Knives	Ceramic Knives	mAP
AlexNet	97.23	99.70	97.30	89.64	93.19	94.50	95.26
GoogLeNet	97.14	92.56	99.50	97.70	95.50	98.40	**98.40**

Table 3: Results for the multi-class problem (average precision %).

4. CONCLUSIONS

This work introduces a technique for the classification of X-ray baggage images using state of the art convolutional neural networks. CNN with transfer learning achieves superior performance compared to prior work [6, 7]. The proposed fine-tuned method achieves 99.26% True Positive (TP) and 95.92% True Negative (TN) with False Positive (FP) and False Negative (FN) rates of 4.08%, 0.74%, respectively. This offers a significant improvement over the prior work [6] which yields TP, TN, FP FN of 85.81%, 88.24%, 11.76%, 14.19% classification. For the classification of multiple X-ray baggage objects, CNN based approaches achieve 95.26% and 98.40% mean average precision rates, clearly demonstrating the applicability of CNN within X-ray baggage imagery.

Future work will consider broader comparison between CNN and hand designed feature descriptors to further investigate the applicability of CNN into this problem domain. Accumulating larger datasets containing various baggage objects will lead to much more realistic scheme for a real time application. Future work will also investigate localization of X-ray baggage objects within the image.

Acknowledgment: Parts of this study were supported by Home Office CAST, Department for Transport, CNPI, Metropolitan Police Service, Defense Science and Technology Laboratory (UK MOD) and Innovate UK.

5. REFERENCES

[1] S. Singh and M. Singh, "Explosives detection systems (eds) for aviation security," *Signal Processing*, vol. 83, no. 1, pp. 31–55, 2003.

[2] M. Mansoor and R. Rajashankari, "Detection of concealed weapons in x-ray images using fuzzy k-nn," *Int. Journal of Comp. Sci., Eng. and Inf. Tech.*, vol. 2, no. 2, 2012.

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape context: A new descriptor for shape matching and object recognition," in *NIPS*, 2000, vol. 2, p. 3.

[4] A. Khotanzad and Y. H. Hong, "Invariant image recognition by zernike moments," *Patt. Anal. and Mach. Intel., IEEE Trans. on*, vol. 12, no. 5, pp. 489–497, 1990.

[5] M. Baştan, M. R. Yousefi, and T. M. Breuel, "Visual words on baggage x-ray images," in *Computer analysis of images and patterns*. Springer, 2011, pp. 360–368.

[6] D. Turcsany, A. Mouton, and T.P. Breckon, "Improving feature-based object recognition for x-ray baggage security screening using primed visualwords," in *Industrial Tech., Int. Conf. on*. IEEE, 2013, pp. 1140–1145.

[7] T. Breuel (Tech. Univ. of Kaiserslautern) M. Bastan (Tech. Univ. of Kaiserslautern), W. Byeon (Tech. Univ. of Kaiserslautern), "Object recognition in multi-view dual energy x-ray images," in *Proceedings of the British Machine Vision Conference*. 2013, BMVA Press.

[8] D. Mery, "X-ray testing by computer vision," in *Computer Vision and Pattern Recognition Workshops*. IEEE, 2013, pp. 360–367.

[9] M. Baştan, "Multi-view object detection in dual-energy x-ray images," *Mach. Vis. and App.*, vol. 26, no. 7-8, pp. 1045–1060, 2015.

[10] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1717–1724.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[13] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.

[14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[15] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. of the 25th Int. Conf. on Machine learning*. ACM, 2008, pp. 160–167.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 580–587.

[17] S. C. Turaga, J. F Murray, V. Jain, F. Roth, M. Helmstaedter, K. Briggman, W. Denk, and H. S. Seung, "Convolutional networks can learn to generate affinity graphs for image segmentation," *Neural Computation*, vol. 22, no. 2, pp. 511–538, 2010.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[20] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," Book in preparation for MIT Press, 2016.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. abs/1311.2901, 2013.

[23] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[24] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[26] A. Mouton and T.P. Breckon, "A review of automated image understanding within 3d baggage computed tomography security screening," *Journal of X-ray science and technology*, vol. 23, no. 5, pp. 531–555, 2015.

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] CC. Chang and CJ. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[29] M. Everingham, L. Van Gool, C KI Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.