IMPROVING FEATURE-BASED OBJECT RECOGNITION FOR X-RAY BAGGAGE SECURITY SCREENING USING PRIMED VISUAL WORDS

Diana Turcsany, Andre Mouton, Toby P. Breckon
School of Engineering, Cranfield University, Bedfordshire, UK.

ABSTRACT

We present a novel Bag-of-Words (BoW) representation scheme for image classification tasks, where the separation of features distinctive of different classes is enforced via class-specific feature-clustering. We investigate the implementation of this approach for the detection of firearms in baggage security X-ray imagery. We implement our novel BoW model using the Speeded-Up Robust Features (SURF) detector and descriptor within a Support Vector Machine (SVM) classifier framework. Experimentation on a large, diverse data set yields a significant improvement in classification performance over previous works with an optimal true positive rate of 99.07% at a false positive rate of 4.31%. Our results indicate that class-specific clustering *primes* the feature space and ultimately simplifies the classification process. We further demonstrate the importance of using diverse, representative data and efficient training and testing procedures. The excellent performance of the classifier is a strong indication of the potential advantages of this technique in threat object detection in security screening settings.

Index Terms— Primed visual words, SURF, BoW, classification, baggage X-ray, airport security

1. INTRODUCTION

Airport security screening personnel are required to manually inspect thousands of items of luggage for contraband on a daily basis. In addition to this enormous workload, X-ray baggage imagery can be extremely challenging to interpret. Due to the nature of packed luggage, where objects are tightly packed, X-ray baggage imagery generally contains a very high degree of clutter. Consequently, objects are often occluded or shown from unusual viewpoints (see Figure 1). It has been shown that both human and computer detection rates are severely affected by complexity and clutter and therefore image interpretation in such environments is particularly challenging [1]. Furthermore, airports are usually overcrowded, demanding high turnover rates at security checkpoints allowing screening personnel only a limited time to examine and classify each item of baggage.

A reliable automated threat detection system for X-ray baggage imagery would significantly speed up the screening process and could improve airport security. The value of such a system extends to 'on-the-job' training and performance evaluation of security personnel. The objective of this study is thus to investigate the application of state-of-the-art object recognition techniques on X-ray images of baggage.

Previous studies which consider the application of computer vision techniques, especially local features-based techniques, are limited in number. Gesick *et al.* [2] evaluate three separate approaches (edge detection combined with pattern matching; an algorithm using Daubechies wavelet transforms [3] and a Scale Invariant Feature Transform (SIFT) [4] based approach) for the detection of weapons in greyscale X-ray imagery. In the case of the first two approaches testing is limited to a very small data set (12 images) yielding unconvincing results. In the case of SIFT, the authors do not evaluate the feasibility of the algorithm for object recognition. Chan *et al.* [5] investigate the use of SIFT for stereo-matching on airport X-ray images. Again, object recognition is not considered, and experimentation is limited to feature matching. Their work demonstrates that while SIFT-based matching can establish correspondences between features, application to X-ray images yields a high ratio of false matches. It is further shown that local intensity patterns in X-ray images of the same content can vary to a greater extent than in corresponding visible spectrum images, making feature matching considerably more challenging. Bastan *et al.* [6] investigate the applicability of the bag-of-words (BoW) model for classification and retrieval of X-ray baggage images. It is shown that a straightforward Support Vector Machine (SVM) classifier using SIFT descriptors yields promising results. However, the authors emphasise performance remains considerably poorer on X-ray baggage images compared to visible spectrum images leaving much room for improvement.

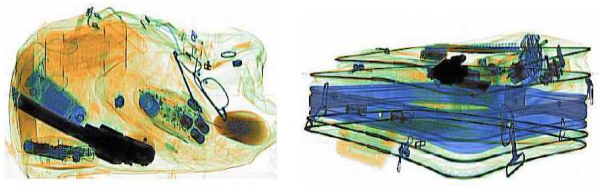


Fig. 1. Two examples of side view X-ray baggage images containing firearms. The images are difficult to interpret due to the unusual viewpoint.

The objective of this study is to develop an automated object recognition system for X-ray baggage imagery, which is able to reliably detect the presence of firearms from both frontal and side-view X-ray images. While we adopt a similar approach to [6] and use the BoW model within an SVM classifier framework, we also propose a novel modification to the traditional codebook generation method. The use of our codebook in constructing the BoW representation of images simplifies the classification procedure by priming the BoW feature space prior to classification. While Bastan *et al.* [6] demonstrate that the use of information rich X-ray imagery (quad-view, dual-energy X-ray images, giving 12 images per item) improves classification performance, we present superior classification results on a dataset of dual-view, single-energy X-ray images (2 images per item). Moreover these results are achieved without the need for using the correspondence between the two different views, instead each image is considered independent during training and testing. We demonstrate the efficacy of our primed BoW approach for object recognition and emphasise the importance of establishing large, diverse data sets and reliable training and testing procedures, thereby extending the previous work of Bastan *et al.* [6].

2. METHODOLOGY

We address the issue of object type recognition as a binary classification problem: image parts which represent a particular target object are distinguished from background parts, which do not contain the target object. In particular, we consider the recognition of firearms

in cluttered 2D X-ray baggage imagery.

The concept of the bag-of-words (BoW) model originated as a document representation technique used in textual information retrieval and text classification. In this original context a document is represented by a frequency vector over words. This simplified representation eliminates all information about the original order of words in the document. In computer vision an image can be represented as a collection of local features, generally in the form of local feature descriptor vectors that encode the local intensity patterns at different image locations. These descriptors are continuous valued multi-dimensional vectors and are therefore of infinite number. Sivic and Zisserman [7] proposed a method to obtain the equivalent of the bag-of-words model for images: local features obtained from an image set are clustered into a finite number of clusters and the cluster centroids form a codebook which is used to encode features of images in a vector quantised representation. The cluster centroids are called visual words and the bag-of-words model represents an image by its histogram over these visual words. In the last few years the BoW approach has been successfully applied to several object recognition and image classification problems [8–12].

Traditionally, image classification using the BoW representation of an image is composed of the following steps [8]: 1) feature detection and description; 2) visual codebook generation; 3) BoW representation and 4) classification. We follow this general framework but also introduce a novel codebook generation technique that significantly simplifies the separation of classes and leads to improved classification results. The details of each of the components of our approach are discussed below.

Feature detection and description: Image representations based on local feature descriptors are widely applied in image classification and object recognition frameworks due to their robustness to partial occlusion and variations in object layout and viewpoint. Distinctive features of objects are detected at interest point locations which generally correspond to local maxima of a saliency measure calculated at each location in an image. The intensity patterns around these interest points are encoded using a descriptor vector. The most widely followed work in the area of local feature extraction has been Lowe's method of the Scale Invariant Feature Transform (SIFT) [4] which introduced a feature descriptor that is invariant to translation, scale and rotation and robust to image noise. Bay *et al.*'s recent work [13] proposed the Speeded Up Robust Features (SURF) algorithm for feature detection and description that is loosely based on SIFT. The computational cost associated with SIFT are dramatically reduced without significant deterioration in performance. This is achieved by introducing box-filter approximations in the calculation of the Hessian matrix-based saliency measure. Integral images are utilised to ensure fast convolution with box-filters and Haar wavelet filters during the detection and description stage respectively. Furthermore according to the authors the SURF algorithm is on par with or even outperforms its counterparts (e.g. SIFT) in terms of repeatability, distinctiveness and robustness of interest point detection and description. Figure 2 illustrates SURF features depicted as circles on baggage X-ray images.

Based on these observations and preliminary experimentation, we employ the SURF method for both detection and description of features. SURF detects interest points corresponding to blobs in images over multiple scales by constructing a scale space. For each identified interest point, the algorithm determines the orientation of the feature and assigns a distinctive, rotation and scale invariant 64-length descriptor vector to the feature. This highly efficient method is robust to noise and changes in 3D viewpoint and illumination [13].

In the context of local feature extraction and object recognition, two important characteristics of X-ray imagery (and baggage imagery in particular) are worth emphasising: 1) objects often appear smeared and lack any informative texture and hence images contain fewer interest points than regular visible spectrum images; 2) X-ray baggage imagery is inherently cluttered which dramatically increases the number of meaningless interest points detected. Considering these two characteristics, it is expected that local feature detection algorithms such as SIFT and SURF are likely to yield excessive interest points corresponding to background clutter and comparatively few interest points corresponding to the target object(s) when applied to X-ray baggage imagery. Two vital steps are taken to address these challenges. Firstly, as a pre-processing step, we perform a coarse foreground segmentation by truncating greyscale pixel intensities above/below a threshold value, prior to applying the SURF detection algorithm. The threshold values are determined empirically such that intensity ranges which do not correspond to the particular target object are excluded. This significantly reduces the number of interest points associated with background clutter. Secondly, in order to increase the number of extracted features on the actual target object, despite the aforementioned lack of textural information, we apply a relatively low threshold value on the saliency measure for interest point identification resulting in a higher number of generated features. Our interest point detection algorithm thereby represents a trade-off between the distinctiveness of features generated by a highly selective interest point detector and the high density of feature locations obtainable from random sampling or applying a dense grid (dense sampling on a regular grid is used by Bastan *et al.* [6]). Although evaluating a saliency measure takes more time than sampling on a regular grid, the SURF detector is still very fast and the reduction in unnecessary (weak) features reduces the required storage and processing time for codebook generation and for forming a BoW representation.

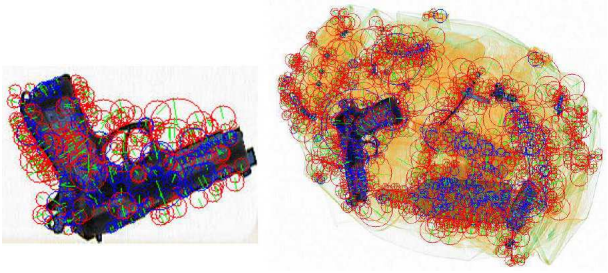


Fig. 2. Detected SURF feature points indicated by circles of varying size (proportional to the scale). Red circles represent light blobs on dark backgrounds; blue circles dark blobs on light backgrounds; green lines indicate dominant orientations of features.

Visual codebook generation: After the feature extraction stage, images of the dataset are represented as varying sized unordered sets of local features. However, most state-of-the-art classification techniques (e.g. SVMs), require the input to be in the form of fixed sized vectors. This problem can be solved by forming a BoW representation of images. The first step in constructing the BoW representation is applying vector quantisation to the feature descriptors. In order to achieve this, a codebook is generated by clustering feature descriptors, usually by a k-means algorithm, then any feature descriptor can be encoded by assigning it to the closest cluster centroid (visual word).

As mentioned, the most popular clustering technique for codebook generation is k-means clustering [14, 15]. In general more accurate clustering can be obtained by the traditional flat k-means method than by hierarchical versions, but for retrieval on large

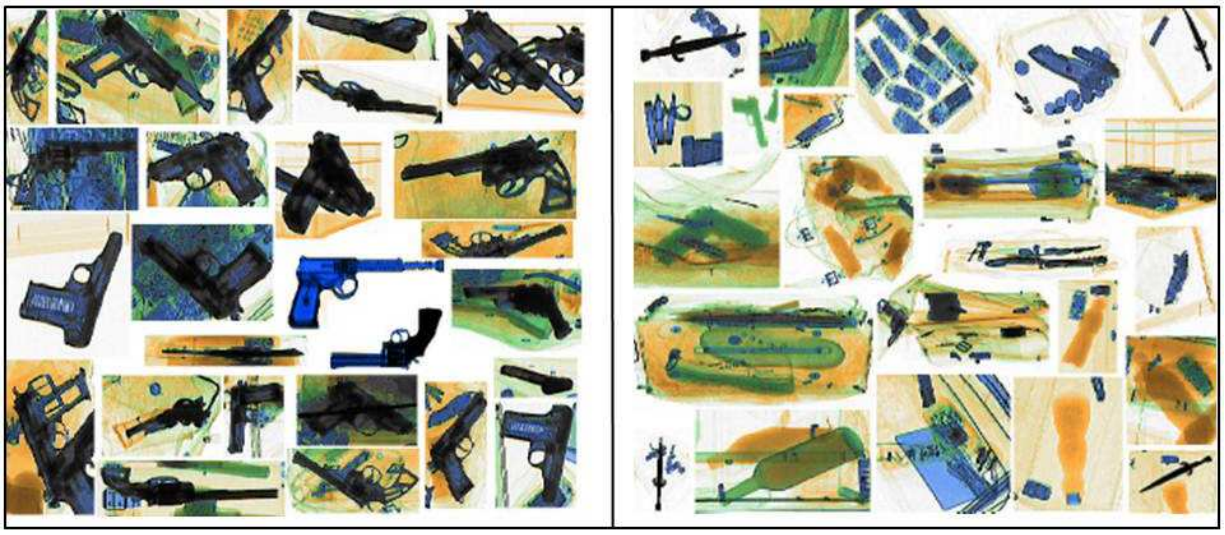


Fig. 3. Training data examples. Positive instances (left) showing firearms of various sizes, shapes, orientations etc. Negative instances (right) showing a variety of clutter items.

datasets it is important to consider fast hierarhical methods [16, 17]. For our classification problem we implement a flat k-means algorithm but choose a variant that reduces memory requirements. As the clustering is to be performed on a high number of samples, an online version of k-means clustering [14] is suitable for our task (see Table 1).

The clustering algorithm initialises the centroids by k distinct feature descriptor vectors randomly selected from the complete feature set. In each iteration, a few features of a randomly chosen image are used for updating the clusters. The Euclidean distance metric is used to determine the closest centroid to a particular feature descriptor. To reduce the computational cost of clustering, only those feature descriptors with the same sign of the Laplacian are compared. The sign of the Laplacian for every feature is the trace of the Hessian matrix at that location which is computed in the SURF descriptor algorithm and distinguishes between dark blobs on light backgrounds and vice versa. As a result, features with different Laplacian signs do not need to be compared [13]. The algorithm terminates when the number of iterations exceeds a predefined limit. This is set to 50,000 in our experiments which we found to be high enough to guarantee that each centroid gets updated enough times.

When generating a visual codebook in a BoW classification framework clustering is performed on the features obtained from a random hold-out subset of the available data, i.e. this subset is not included during classifier training and testing. Traditionally, one run of the clustering method is performed using all images from this subset regardless of the class of the object they represent, resulting in a single set of centroids, the codebook. It is then the task of the classifier to identify class specific patterns in the visual word histograms [15].

We propose an intuitive modification to the aforementioned clustering technique which we believe dramatically simplifies the classification procedure into different object classes leading to improved results. Our clustering method is also performed on features of a 'hold-out' image set but we do not ignore class labels. Rather, we perform clustering independently on each class resulting in multiple sets of class-specific centroids (visual words). A similar technique has been met with considerable success by Perronnin *et al.* [18].

Cluster centroids: $c_1, \ldots, c_k$

Number of updates per cluster: $n_1, \ldots, n_k$

Initialise cluster centroids by randomly selected data points

$n_i = 1$ $(i = 1, \ldots, k)$

while $iteration\_count < max\_iteration\_count$

select random data point: x

find nearest centroid to x

$t = argmin_{i=1,\ldots,k} \|x - c_i\|$

update cluster centroid c_t

$n_t = n_t + 1$

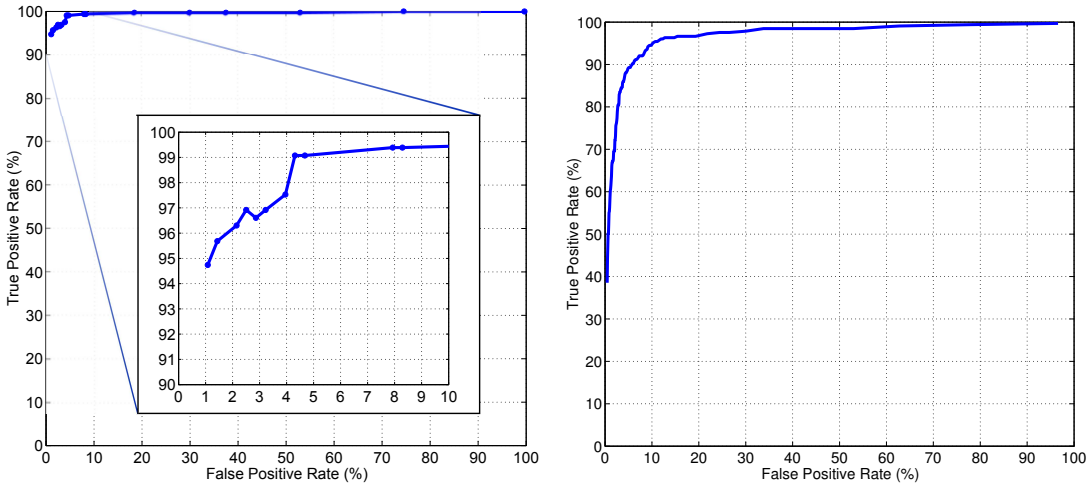
$c_t = c_t + \frac{x - c_t}{n_t}$

end

Table 1. Online k-means clustering method.

In our experiments we categorised X-ray images into two classes: *positive* which represent the target object and *negative* which represent background i.e. all 'target-free' images (see examples of positive/negative images for firearm classification in Figure 3). Therefore in the followings, the description of the clustering method is tailored to a two-class classification problem, but we emphasise here that the method is also suitable for multi-class classification.

After performing feature clustering on each class, the obtained sets of centroids are combined to form a codebook. The process of encoding image features using this codebook is the same as in a traditional BoW model. The motivation behind the isolated per-class clustering approach is to enforce a representational separation of the positive and negative class examples in the BoW model. This *primed* BoW representation significantly simplifies the task of the classifier. An added advantage to performing separate clustering on positive and negative classes is the ability to influence the ratio of visual words representative of each class in the codebook. In many classification datasets negative examples are overrepresented (as in our case) because usually they are easier to obtain. One would like to utilise a diverse set of negative examples during codebook gen-



(a) Firearm classification (b) Mobile classification

Fig. 4. ROC curve for classification of (a) firearm and (b) mobile phone images. In both cases the free parameter is the RBF kernel width σ. In order to facilitate examination, (a) also contains a graph that is zoomed in on high true positive values.

eration along with guaranteeing that the distinctive features of the positive examples are well represented in the codebook. Our feature clustering method provides a simple solution to this problem as one can determine beforehand how many visual words are obtained during each class-specific run of the clustering method. In our experiments we generate equal amounts of visual words (k) from both positive and negative classes which results in a $2k$-sized codebook.

Bag-of-words representation: Up to this point images have been represented by their collections of local features. Once the visual codebook has been generated, this image representation can be transformed into a fixed size vector. To this end, each feature descriptor is encoded by hard assignment to the cluster it belongs to, which is given by the nearest visual word in the codebook according to Euclidean distance. This vector quantisation of features is not only important for obtaining suitable image representation for classification but also reduces noise due to minor differences in the descriptor vectors of corresponding features. By assigning each feature of an image to the appropriate visual word and accumulating the word-counts one can obtain a histogram over visual words (bag-of-words). This histogram gives a highly generalized representation of the image content due to its inherent robustness to noise and changes in scale, rotation and viewpoint. Since our codebook consist of $2k$ visual words, the transformation yields a $2k$-dimensional feature vector per image. The image features are now represented in a form which allows for integration into any common classifier algorithm.

Classification: Support Vector Machines (SVMs) [19] are regarded as one of the most powerful classification tools. SVMs attempt to determine an optimal linear separation of classes by maximising the margin of separation between classes. Using this criterion, optimisation results in a separator that can be recovered at any time using only a few data points: namely those lying nearest to the boundary of separation (and hence determining the margin). These data points are aptly called support vectors and can be used to identify the class of a new observation. While the described method fails in cases of linearly inseparable data, separation is still attainable via a higher-dimensional hyperplane. To be able to distinguish between classes by a maximal margin separator, the data points are projected into a higher dimensional space using a suitable predefined non-linear kernel function. The most suitable kernel type and optimal parameter settings for this study were determined empirically. In our initial experimentation, the Gaussian Radial Bases Function (RBF) kernel [20] yielded the optimal classification results and is thus used in the remainder of this work. The RBF kernel function is defined using the following formula:

$$K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \quad (1)$$

where σ is the kernel width. Intuitively, the centre of the RBF represents the support vector, while σ determines the area of influence that this support vector will have over the data space. Increasing σ increases the neighbourhood of influence of the support vector, resulting in smoother, more regular decision boundaries. The optimal choice of σ is data dependent and one runs the risk of overfitting if σ is chosen to be too small [20]. The performance of an SVM classifier is heavily influenced by both the choice of kernel as well as the kernel parameters [19]. We thus measure the performance of the SVM classifier over a range of values for σ and represent the results on a Receiver Operating Characteristic (ROC) curve [21].

3. RESULTS

The aforementioned techniques were tested on a dataset of 2500 pseudo-coloured X-ray baggage images obtained by scanning packed handbags with a dual-view, single-energy Rapiscan 620DV scanner. In these pseudo-coloured X-ray images darker pixels indicate areas where the density of materials is high (see Figure 1). Two images (one frontal and one side view image) were generated for every scanned item. The bags contained various types of threat objects including handguns, knives, explosives and bottles filled with fluids of varying densities as well as everyday items representing clutter (see Figure 3).

For the purposes of this work, we considered the specific problem of firearm recognition. A total of 850 firearm images were manually cropped from both the frontal and side-view scans and used as positive instances. Importantly, the positive training data captured a large range of variation in the type, size, orientation and degree of occlusion of the firearm. Approximately 10 000 images without

	This study	Bastan *et al.* [6]
Pre-processing	foreground segmentation	foreground segmentation
Interest point detection	SURF	DoG + Harris
Descriptor	SURF	SIFT
Codebook generation	class-specific online k-means clustering	traditional k-means
Size of codebook (number of clusters)	1200	200
Vector quantisation	Hard assignment	Soft assignment
Classifier	SVM	SVM
Kernel	Gaussian RBF	Histogram intersection
Experimentation	3-fold cross validation	Separate training and test sets
Scanner	Dual-view, single energy X-ray (2 images per item)	Quad-view dual-energy X-ray (12 images per item)
Data	850 pos; 10000 neg	208 training (52 pos; 156 neg) 764 test (40 pos; 724 neg)

Table 2. Comparison of proposed approach and method of Bastan *et al.* [6]

firearms were automatically generated from randomly cropped regions of baggage images (with some added hand-cropped regions) and used as negative examples. Figure 3 shows several instances of positive and negative training data. To avoid overfitting the data, these images were used in a 3-fold cross validation framework. We ensured that different X-ray images of the same packed baggage were not included in both training and testing sets.

The SURF method was applied for feature detection and description, Figure 2 shows two example images where rotation and scale invariant SURF features have been detected on two thresholded images. Online k-means clustering of features with $k = 600$ was performed separately on the firearm and background images, generating two class-specific sets of visual words containing 600 words each, which gives a total of 1200 visual words in the codebook.

Performance evaluation was performed using the standard Receiver Operating Characteristic (ROC) curve [21] which plots the True Positive Rate (TPR) as a function of the False Positive Rate (FPR) as the free parameter (RBF kernel width σ) is varied. The ROC curve in Figure 4(a) shows the results of the 3-fold cross validation testing. The optimal operating point yielded a correct detection rate of 99.07% at a false positive rate of 4.31%. In our threat item detection application high true positive rates are favoured, therefore when reporting an optimal operating point we select a value from the ROC curve which favours a high TPR. The excellent results of the classifier indicate that the distinctive image features have been generalised successfully by the clustering algorithm and the SVM classifier established a good separation of classes.

We also show preliminary results for the detection of mobile phones using the same method outlined for firearms. A smaller dataset of 350 mobile images and 1700 background images were used in a 3-fold cross-validation. Mobile phones are tiny in size compared to firearms and therefore tend to contain less features. Also, their colour is prevalent among objects in the background clutter causing thresholding to be less effective in background removal. Despite these difficulties we still obtained very good results for mobile phone detection and show the ROC curve in Figure 4(b), obtained again by altering the RBF kernel width σ.

4. DISCUSSION AND RELATED WORK

The most similar method to our own within previously published literature on threat item detection in airport baggage is Bastan *et al.*'s system [6]. While the performance analysis tools differ, the aforementioned result (TPR = 99.07%) is a significant improvement on the optimal results (TPR = 70%) presented in their work. While both studies employ a BoW approach using an SVM classifier, the two approaches differ in their particulars (see Table 2). The first significant differences occur in the feature detection (SURF vs. a combination of DoG and Harris) and description (SURF vs. SIFT) techniques and the choice of the SVM kernel (RBF vs. intersection). While it is accepted that variations in these components may have a significant impact on the overall performance, Bastan *et al.* [6] indicate that their choice of components is based on prior experimentation, which ultimately ruled out SURF descriptors and the RBF kernel. It is therefore unlikely that our improved classification results stem from our choice of feature representation and/or SVM kernel.

We thus believe that the core of our improved classification results lies in our novel codebook priming technique achieved by class-specific clustering. By enforcing the separation of classes in the bag-of-words representation scheme, the classification process is dramatically simplified. This priming, coupled with the use of a larger codebook (1200 visual words vs. 200 visual words), ultimately leads to superior results. In general larger codebooks yield better classification results [8, 17].

The final significant difference is data-related. Although the images used by Bastan *et al.* [6] (quad-view, dual-energy X-rays) should theoretically provide more useful information than the images used in our work (dual-view, single-energy X-rays), we use a much larger data set ($\sim$11,000 images vs. $\sim$1000 images). Our results are also possibly more reliable than the result published by Bastan *et al.* [6] as we employ 3-fold cross-validation as opposed to using a training and a test set. The importance of the k-fold cross-validation approach in performance evaluation is widely known [22]. We believe that the diversity and richness of our data, as well as more efficient use of this data has further contributed to our superior results.

While Bastan *et al.* [6] claim that their initial experimentation indicated inferior performance for SURF features and an RBF kernel SVM classifier, we have shown that even superior results can be obtained using these components in a similar classification framework by making more efficient use of the available data and priming the BoW representation in order to simplify classification. We also presented good classification results on mobile phones. We expect that using X-ray imagery capable of capturing more images per item from different viewpoints, as used by Bastan *et al.* [6], will further improve our results. Our experiments indicate that local feature-based classification techniques are powerful tools for the recognition of firearms and other objects in X-ray images.

Computational costs: Given that our ultimate aim is to in-

troduce automated object recognition methods such as the one described in this study to the domain of airport security, it is important to ensure the feasibility of our proposed methods for real time execution. With this in mind, our choice of algorithm placed preference on efficient methods provided that accuracy was not compromised. To reduce any unnecessary computational cost we employed a faster alternative to the SIFT feature detector: the SURF algorithm. Furthermore, as opposed to the traditional offline k-means, we implemented an online k-means clustering algorithm. This algorithm reduces the required memory and is suitable for clustering large datasets. Finally, when feature descriptor encoding was performed using a hard assignment of features to clusters as opposed to a soft assignment thereby keeping the computational costs to a minimum.

Importantly, the most computationally expensive components of our method (clustering and classifier training) are done offline. During testing, when a new image is presented to the system, 3 subunits have to be executed: 1) detecting local features in the image; 2) constructing the BoW representation via a nearest neighbour lookup of each feature descriptor in the codebook and 3) feeding the BoW histogram to the pre-trained SVM classifier. With an optimised implementation, these units can all be executed in real time and numerous systems have applied these or similar methods successfully even for large-scale problems [7, 17, 23]. Calculating the SURF features of an image of size 800x640 takes approximately 0.61s [13]. Our C++ implementations performed on a standard PC (Intel Core 2 1.83GHz CPU with 1 GB RAM) yielded processing times of 1.5s per image for SURF feature calculations; 0.03s per image for the BoW construction and 0.003s per image for the SVM prediction.

5. CONCLUSION

This study has presented a powerful image classification technique for object detection in X-ray baggage imagery using primed visual words in an SVM classifier framework. Primed visual words are obtained through class-specific clustering of feature descriptors and used to encode images in our bag-of-words model. This differs from the traditional approach, which combines the feature set of positive and negative classes during the clustering process when generating a codebook. Our novel modification to the clustering stage of the traditional bag-of-words framework creates an image representation scheme that further facilitates the separation of positive and negative class examples.

The proposed method has been evaluated on a firearm recognition problem and yielded excellent results with an optimal operating point of 99.07% TPR and 4.31% FPR on the ROC curve. The method presented here thus significantly outperforms the previous work of Bastan *et al.* [6]. We have also shown promising results in the more challenging task of detecting mobile phones. Finally, we have demonstrated the value of establishing large, representative data sets for improving classification results. The excellent performance of our firearm classifier indicates the potential of this technique in threat object detection.

Future work will consider the classification of additional object types, the use of different X-ray imagery (capturing more images per item) and an investigation into further optimisation techniques to facilitate the real time application of the proposed method.

6. REFERENCES

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

[2] R. Gesick, C. Saritac, and C.C. Hung, "Automatic image analysis process for the detection of concealed weapons," in *Proceedings of the Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies*, 2009, pp. 20.1–20.4.

[3] I. Daubechies, *Ten lectures on wavelets*, Society for Industrial and Applied Mathematics, 1992.

[4] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] J.W. Chan, A. Omar, J.P.O. Evans, D. Downes, X. Wang, and Y. Liu, "Feasibility of SIFT to synthesise KDEX imagery for aviation luggage security screening," in *International Conference on Crime Detection and Prevention*, 2009, pp. 1–6.

[6] M. Bastan, M. Yousefi, and T. Breuel, "Visual words on baggage X-ray images," in *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, 2011, pp. 360–368.

[7] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, 2003, pp. 1470–1477.

[8] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proceedings of the European Conference on Computer Vision*, 2006, pp. 490–503.

[9] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, vol. 2, pp. 264–271.

[10] E.B. Sudderth, A. Torralba, W.T. Freeman, and A.S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *IEEE International Conference on Computer Vision*, 2005, vol. 2, pp. 1331–1338.

[11] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[12] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[14] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 1, pp. 281–297.

[15] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proceedings of the British Machine Vision Conference*, 2011, pp. 76.1–76.12.

[16] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2161–2168.

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[18] F. Perronnin, C. Dance, G. Csurka, and M. Bressan, "Adapted vocabularies for generic visual categorization," in *Proceedings of the European Conference on Computer Vision*, 2006, pp. 464–475.

[19] V.N. Vapnik, *The nature of statistical learning theory*, Springer, 2000.

[20] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Methods in Molecular Biology*, vol. 609, pp. 223–239, 2010.

[21] C.E. Metz, "Basic principles of ROC analysis," in *Seminars in Nuclear Medicine*, 1978, vol. 8, pp. 283–298.

[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–1145.

[23] Y. Li, D.J. Crandall, and D.P. Huttenlocher, "Landmark classification in large-scale image collections," in *IEEE International Conference on Computer Vision*, 2009, pp. 1957–1964.