# On using Feature Descriptors as Visual Words for Object Detection within X-ray Baggage Security Screening

**Mikolaj E. Kundegorski**[1], **Samet Akçay**[1], **Michael Devereux**[2], **Andre Mouton**[3], **Toby P. Breckon**[1]
[1]Durham University, UK [2]Heriot-Watt University, UK [3]University of Bath, UK

**Keywords:** X-ray security screening, automatic threat detection, firearms detection, bag of visual words

## Abstract

Here we explore the use of various feature point descriptors as visual word variants within a Bag-of-Visual-Words (BoVW) representation scheme for image classification based threat detection within baggage security X-ray imagery. Using a classical BoVW model with a range of feature point detectors and descriptors, supported by both Support Vector Machine (SVM) and Random Forest classification, we illustrate the current performance capability of approaches following this image classification paradigm over a large X-ray baggage imagery data set. An optimal statistical accuracy of 0.94 (true positive: 83%; false positive: 3.3%) is achieved using a FAST-SURF feature detector and descriptor combination for a firearms detection task. Our results indicate comparative levels of performance for BoVW based approaches for this task over extensive variations in feature detector, feature descriptor, vocabulary size and final classification approach. We further demonstrate a by-product of such approaches in using feature point density as a simple measure of image complexity available as an integral part of the overall classification pipeline. The performance achieved characterises the potential for BoVW based approaches for threat object detection within the future automation of X-ray security screening against other contemporary approaches in the field.

## 1 Introduction

Within transport security, screening personnel are required to manually inspect thousands of baggage items for a range of 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 baggage, where objects are tightly packed, X-ray imagery generally contains a very high degree of clutter and inter-object occlusion. Consequently, objects are most 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. Furthermore, increasing global travel demands ever increasing turnover rates at security checkpoints allowing screening personnel only limited time to examine each baggage item.

A reliable automated threat detection system for X-ray baggage imagery that can automatically detect the presence of threat item characteristics offers the potential to significantly stream-line this screening process and facilitate an extended threat screening footprint beyond the conventional remit of passenger carry-on baggage. This poses an interesting challenge for the use of automatic object recognition approaches akin to the prior work of [1, 2, 3, 4]. In addition, an associated ability to automatically assess the underlying complexity of a given X-ray baggage image facilitates the potential of "auto-clearing" low complexity baggage items (e.g. comprised solely of paperwork, clothing) and similarly "auto-referring" high complexity items that maybe more challenging for automatic detection approaches.

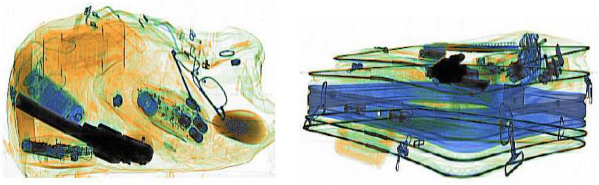

Figure 1. Typical X-ray baggage imagery.

In general, prior work on object detection in X-ray baggage imagery is limited. Inspired by implicit shape models, Mery [5] proposes a method that automatically detects X-ray baggage objects using a visual vocabulary concept, occurrence structures with $99\%$ and $0.2\%$ true positive and false positive achieved for handgun detection over 200 example images. Shape-based handgun detection is further investigated in [6] by training fuzzy k-NN classifier but with limited evaluation over only 15 image examples. The work of Baştan et al. [1] considers the concept of Bag-of-Visual-Words (BoVW) within X-ray baggage imagery using Support Vector Machine (SVM) classification with SIFT feature descriptors [7] achieving performance of $0.7$, $0.29$, $0.57$ recall, precision and average precision, respectively. Turcsany et al. [4] followed a similar approach, extending the work of [1], using BoVW with SURF feature descriptors [8] and SVM classification together with a modified version of vocabulary generation to yield $99.07\%$ true positive, and $4.31\%$ false positive on firearms detection over 2000 examples. A BoVW approach with SIFT feature descriptors, augmented with SPIN X-ray intensity features [9], and SVM classification is also used in [3] for the classification of single and dual view X-ray images with best average precisions achieved for gun and laptop objects of $94.6\%$ and $98.2\%$. Baştan thoroughly reviews several feature detectors (Harris–Laplace, Harris-affine, Hessian–Laplace, Hessian-affine) in his latest work [2], on which he studies applicability and efficiency of sparse local features (SIFT + SPIN [7, 9]) on object detection in X-ray baggage imagery via the use of a similar Bag-of-Features concept. This work also investigates how material information given in X-ray imagery via colour mapping (Figure 1) and multi-view X-ray imaging affect detection performance [3, 2].

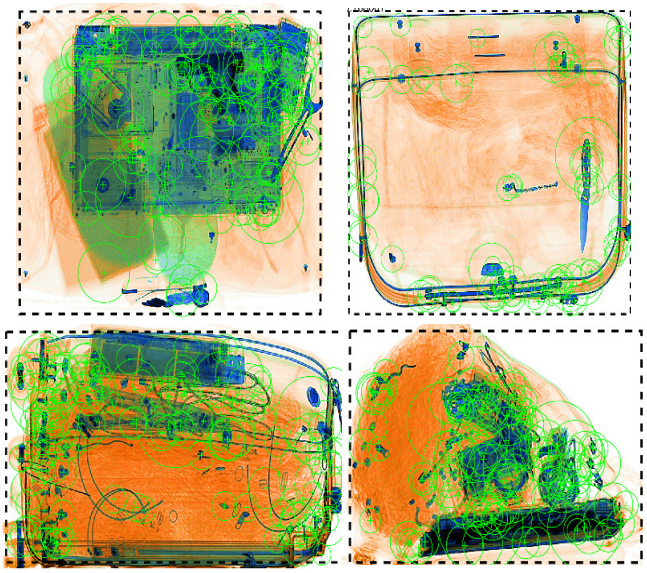

Figure 2. SURF feature points (green) give a raw indication of composition complexity across the X-ray image.

A related body of work in 3D Computed Tomography (CT) for baggage security has investigated a 3D BoVW approach with suitable extensions to the relevant feature detection and descriptor approaches [10, 11, 12, 13]. This work largely concludes that the choice of the feature descriptor, feature sampling/detection strategy and the final classification framework had a significant impact on performance [10, 11, 13] with the use of simplistic feature descriptors notably outperforming 3D derivatives of established approaches (e.g. SIFT / RIFT [10]).

With a notable common focus across a range of prior work in this domain on the BoVW feature representation approach [1, 4, 5, 3, 12, 11, 2], the objective of this study is thus to investigate the relative performance of a number of state-of-the-art feature point detection and description approaches for this task. With contrasting trends towards the use of end-to-end convolutional neural network (CNN) for the complete feature detection to classification pipeline (e.g. [14, 15]), here we aim to provide a definitive benchmark of over a range of hand-tuned features of varying complexity (namely: FREAK [16], DAISY [17], BRISK [18], ORB [19], KAZE [20], AKAZE [21]) against the mainstay of prior work in the field (i.e. SIFT [7] / SURF [8] with [1, 4, 3, 2]). To these ends, we present an overview of a classical BoVW architecture (Section 2) into which we present both the by-product availability of complexity analysis (Figure 2) and our comparative performance evaluation (Section 3).

## 2 Bag-of-Visual-Words

Following the prior work of [4], we address the issue of automatic threat 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 detection in cluttered 2D X-ray baggage imagery using a feature-driven approach known as Bag-of-Visual-Words (BoVW) (or simply bag-of-words (BoW) in earlier work, [4]).

The concept of the BoW model has origins as a document representation technique used in text information retrieval and text classification. Within this original context a document is represented by a simple frequency vector of words occurrence eliminating all information about word order. In an image classification context, 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 varying image locations. These descriptors are continuous valued multi-dimensional vectors and occurring at various points of localized salience within the image. Sivic et al., [22] originally 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 (vocabulary) which is used to encode features of images in a vector quantized representation. These cluster centroids are called visual words and hence the BoVW model now represents an image as a histogram of visual words occurrence. Following widespread uptake for generalized object recognition, recent work in X-ray imagery has followed this paradigm [4, 12].

Traditionally, image classification using the BoVW representation of an image is composed of the following stages: 1) feature detection and description; 2) visual codebook generation; 3) BoW representation and 4) classification. We follow this general framework introducing variation in the core feature detection and description (descriptor) stage. The details of each of the components of this 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) [7] which introduced a feature descriptor that is invariant to translation, scale and rotation and robust to image noise (as used in X-ray object detection work of [1, 3, 2]). Bay *et al.*'s later work [8] 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 (as used in X-ray object detection work of [4] and later for comparison in the CNN work of [14]).

More recently, research in this area led to industrious efforts to optimize sparse feature stability against computational performance leading to a range of local feature and detector variants. A standalone feature detector FAST (Features from Accelerated Segment Test) [23] provides significant number of candidate points for extraction while maintaining low computational cost. The detector-extractor frameworks BRIEF (Binary Robust Independent Elementary Features) [24], and BRISK (Binary Robust Invariant Scalable Key-points)[18] offer integer-space representations, avoiding the floating point operation of earlier SURF/SIFT variants, for faster extrac-

tion and subsequent computation on embedded platforms. ORB[19] (Oriented FAST and Rotated BRIEF) extends such methods to address issues of rotation invariance. A recent pairing of floating-point and integer space feature frameworks KAZE [20] and AKAZE [21] aim to improve feature uniqueness and robustness of features by describing them based on a non-linear model of an image. More recently FREAK (Fast Retina Key-point) [16], following from the earlier DAISY [17], represent feature extractors specifically inspired by retinal sampling in the human visual system.

Although many further variants exist, here we identify a board range of such feature descriptors which are subsequently evaluated both with their original and variant initial feature point detection approaches (Tables 2 / 3). In all cases, the density of these locally-salient feature points gives rise to both a local and global means of measuring image complexity as a by-product of the BoVW process (e.g. Figure 2).

**Visual codebook generation:** After the feature extraction stage, a given image is now represented as a variable size set of unordered local features (e.g. Figure 2). However, most state-of-the-art classification techniques (e.g. SVM / Random Forest) require a fixed dimensionality of vector input. This problem is essentially solved using the BoVW feature representation. The first step is to apply 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 via fast approximate nearest neighbour matching [25], such that any feature descriptor can be subsequently encoded by assigning it to the closest cluster centroid (visual word) within the resulting set, $k_v$.

**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 dimension feature 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 either Euclidean distance (for floating point descriptors) or Hamming distance (for binary descriptors). This vector quantization 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 (BoVW). 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. The image features are now represented in a form which allows for integration into any common classification algorithm.

**Classification:** From this BoVW feature encoding of feature descriptors, we have an overall feature representation of dimension $k_v$ (the number of visual code words used in our earlier bag of visual words vocabulary/codebook). SVM [26] and Random Forest (RF) [27] classifiers are trained using this encoded feature representations over a corpus of exemplar imagery (X-ray image patches, Figure 3). SVM are trained using

Table 1. Mean execution time and feature density.

| Detector | Descriptor | density (%) | execution (ms.) |
|---|---|---|---|
| SURF [8] | SURF [8] | 0.24 | 4.7 |
| (Hessian = 100, octaves = 4, octave layers = 3, dim = 64) | | | |
| SIFT [7] | SIFT [7] | 0.21 | 12.7 |
| (octaves = 3, contrast = 0.04, edge = 10, $\sigma$ = 1.6) | | | |
| ORB [19] | ORB [19] | 0.66 | **1.4** |
| (scale = 1.2, levels = 8, patch size = 5, threshold = 5) | | | |
| KAZE [20] | KAZE [20] | 0.19 | 17 |
| (threshold = 0.001, octaves = 4, layers = 4) | | | |
| FAST[23] | SURF [8] | 1.22 | 5.8 |
| (threshold = 0, with non-maximal suppression, as above) | | | |
| FAST[23] | SIFT [7] | 1.22 | 51.4 |
| (as above) | | | |
| FAST[23] | ORB [19] | 1.13 | **1.1** |
| (as above) | | | |
| FAST[23] | FREAK [16] | 1.22 | 4.1 |
| (as above + octaves = 4, scale = 0.1) | | | |
| FAST[23] | DAISY [17] | 1.22 | 8 |
| (as above + radius = 15, $q_{radius} = 3, q_\theta = 8, q_{hist} = 8$)) | | | |
| FAST[23] | BRISK [18] | 1.22 | 4.2 |
| (as above + scale = 0.1) | | | |
| BRISK [18] | BRISK [18] | 1.70 | 14 |
| ( threshold = 10, octaves = 3, scale = 0.1) | | | |
| AKAZE [21] | AKAZE [21] | 0.04 | 3.2 |
| (threshold = 0.001, octaves = 4, layers = 4) | | | |

Radial Basis Function (RBF) kernel $\{SVM_{RBF}\}$ with a grid search over kernel parameter, $\gamma = 2^x : x \in \{-15, 3\}$, and model fitting cost, $c = 2^x : x \in \{5, 15\}$, using $k-$fold cross validation ($k = 5$)., using $k-$fold cross validation ($k = 5$) with F-score optimisation (being more representative then accuracy for unbalanced data set). RF are trained over varying values of maximal tree depth, $d = \{5..50\}$, with maximal number of trees per forest, $t = \{1000, 5000\}$ and minimal sample count per leaf node set to equal 1% of training data. The results for the best performing parameter set are reported for each feature configuration (Section 3).

## 3 Evaluation

Our evaluation datasets consist of 19398 X-ray sample patches (dual-energy, false-coloured, from varying manufacturers) upon which we evaluate our feature point detector and descriptor approaches on a classical two-class firearms detection problem (positive class: 3179 gun images / 1176 images of gun components; negative class: 476 images of cameras, 2750 knives, 1561 ceramic knives, 995 laptops and 9261 cropped images of background clutter). This is randomly split into training (67%) and validation test subsets (33%) with the former used for $k$-fold cross-validation based training ($k = 5$, see examples in Figure 3). The parameters used for the feature detectors and descriptor combinations evaluated are listed in Table 1 together with average feature detection density (as percentage of image resolution) and execution time (measured as CPU time used in milliseconds, ms) calculated over a random subset of 1400 X-ray image patches taken from the training set (resolution: $256 \times 256$ pixels).

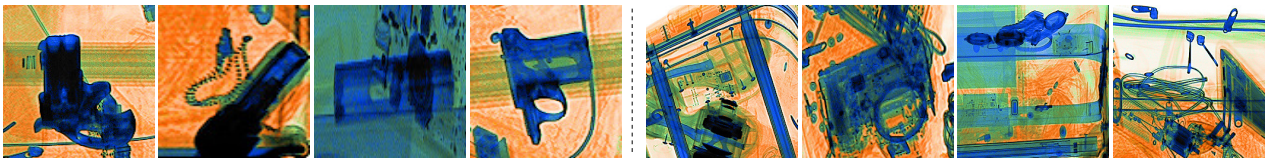

Figure 3. Training data examples. Positive examples (left) showing firearms and sub-components of various sizes, shapes, orientations etc. Negative examples (right) showing a variety of clutter items.

Within the feature detector, descriptor and classification variants outlined, we consider the comparison of True Positive Rate (TP), False Positive Rate (FP) (%) together with the Precision (P), accuracy (A) and F-score (F) (harmonic mean of precision and true positive rate) as shown in Tables 2 & 3.

From these results (Tables 2 and 3) we can see that the best performance was achieved using a FAST-SURF feature detector and descriptor combination with a $k_v = 2000$ BoVW vocabulary and SVM based classification (A: 0.94, TP: 83%, FP: 3.3% - Table 2). This optimal performance is closely followed by SURF-SURF (A: 0.93), FAST-SIFT and KAZE-KAZE (both A: 0.92) feature detector and descriptor combinations (using $k_v = 1500$ vocabulary and SVM, Table 2). These four feature detector and descriptor combinations notable outperform other approaches with F-scores of 0.85 (FAST-SURF), 0.84 (SURF-SURF), 0.81 (FAST-SIFT) and 0.81 (KAZE-KAZE) (for $k_v = 1500$) that are significantly higher than the next best (fifth ranking) method SIFT-SIFT (F-score 0.75, $k_v = 2000$). The FP is just over 3% for the overall best performing classifiers, dropping to 1.9%v for Random Forest trained KAZE-KAZE at expense of significantly lower TP of 54.9% (Table 3). Vocabulary size has only a marginal difference towards the final result suggesting that vocabulary $k_v = 500$ is enough to create a viable feature model and the final result is limited by discriminative properties of the feature detector and descriptor combination in use. Overall the SVM results (Table 2) are consistently better than the RF results (Table 3) for all feature detector and descriptor combinations.

From Table 1 we can see that by contrast the FAST-ORB feature detector and descriptor combination gives the best computational efficiency (under 1.5ms) compared to the others. However the slower, high accuracy combinations (from Table 2), FAST-SURF (5.8ms) and SURF-SURF (4.7ms), significantly outperform the next best performing combinations (FAST-SIFT and KAZE-KAZE) in terms of efficiency.

Overall the best detection accuracy is achieved using variants of the SURF [8] feature detector and descriptor which also perform very computationally efficiently within this side-by-side comparison. This supports the prior work in the field using both SURF [4] and SIFT [3, 1, 2] based BoVW variants for X-ray object detection tasks and comprehensively shows comparative statistical accuracy and relative computational efficiency over a common dataset of ∼20,000 examples. The strong performance of the simpler SURF [8] feature detector and descriptor in place of more complex approaches echo analogous findings in the CT-based object detection literature where simpler density-based descriptors were notably found to outperform contemporary 3D extensions of the seminal SIFT approach [10, 12].

In addition to this statistical analysis, exemplar detection results for both whole firearms and firearm components in cluttered and challenging X-ray imagery are shown in Figures 4 (A-G) and 5 (A-E) where we see both the detected item highlighted (Figures 4 / 5, left - red box) and the resulting heat map of localized detection strength (Figures 4 / 5, right). Based on the BoVW classification approach outlined, object localization within the X-ray image is performed using a classical sliding window search strategy where-by each sub-region (patch) of the X-ray image is individually classified for the presence/absence of the target object. These are then aggregated based on classification response which is based either on normalized distance from the SVM classification boundary or from the majority vote of the RF ensemble classification. This scanning window approach gives rise to both the heat map of localized detection strength (left) and the aggregated minimal object bounding box (right) as shown in Figures 4 and 5. The dimension of the scan window is of fixed size, $w_x \times w_y$, determined from the scale of the scan plane to image plane projection used by the X-ray scanner in use (here: $w_x \times w_y = \frac{image\ width}{5} \times \frac{image\ height}{5}$; with step-size= 60). Overlapping or non-overlapping scan windows can be employed, however, the structureless nature of the BoVW descriptor that makes it notably robust to object occlusion [22] and similarly supports the detection of object sub-parts split between multiple window patches (see examples of Figure 3) as used here (Figures 4 / 5). Within this work, explicit multiple scale object detection is not performed due to the parallel (scale preserving) nature of the X-ray scanner image plane projection and the inherent scale invariance of the BoVW approach [22]. If the scanner projection is not scale preserving for some reason (e.g. images are non-uniformly re-scaled for regular human review) then an alternative multi-scale classifier training and subsequent scan window search strategy maybe required.

## 4 Conclusions

This work has re-enforced the capability of BoVW techniques for object detection in X-ray imagery providing a comprehensive comparison of feature detector and descriptor approaches on the sample task of firearm detection under varying vocabulary sizes and classification approaches. It shows optimal performance of a combined FAST [23] feature detector and SURF [8] feature descriptor over other contemporary approaches (A: 0.94, TP: 83%, FP: 3.3%) over a significantly larger data set than in previous work [1, 5, 3, 4, 2]. The results supports the choice and performance of this descriptor in early isolated studies [4] and as a contemporary comparator for BoVW against other techniques (e.g. [14]). An object detection capability for firearms, and an implicit ability to offer image complexity

Table 2. Results of feature point detector and descriptor variants using BoVW with $SVM_{RBF}$ classification for firearm detection.

| | | $k_v = 500$ | | | | | $k_v = 1000$ | | | | | $k_v = 1500$ | | | | | $k_v = 2000$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Detector | Descriptor | TP | FP | P | A | F | TP | FP | P | A | F | TP | FP | P | A | F | TP | FP | P | A | F |
| SURF [8] | SURF [8] | **80.8** | **4.0** | **0.85** | **0.93** | **0.83** | **79.8** | **3.6** | **0.87** | **0.93** | **0.83** | **79.3** | **3.1** | **0.88** | **0.93** | **0.84** | **79.2** | **3.2** | **0.88** | **0.93** | **0.83** |
| SIFT [7] | SIFT [7] | 70.1 | 5.7 | 0.78 | 0.89 | 0.74 | 68.3 | 4.5 | 0.82 | 0.89 | 0.74 | 69.1 | 4.9 | 0.81 | 0.89 | 0.74 | 68.3 | 4.2 | 0.83 | 0.90 | 0.75 |
| ORB [19] | ORB [19] | 85.6 | 27.0 | 0.48 | 0.76 | 0.62 | 81.2 | 26.4 | 0.47 | 0.75 | 0.60 | 80.2 | 24.1 | 0.49 | 0.77 | 0.61 | 82.3 | 23.6 | 0.51 | 0.78 | 0.63 |
| KAZE [20] | KAZE [20] | **78.3** | **4.6** | **0.83** | **0.92** | **0.81** | **78.6** | **3.6** | **0.86** | **0.92** | **0.82** | **76.9** | **4.1** | **0.85** | **0.92** | **0.81** | **77.3** | **3.9** | **0.85** | **0.92** | **0.81** |
| FAST [23] | SURF [8] | **81.4** | **3.5** | **0.87** | **0.93** | **0.84** | **81.4** | **3.3** | **0.88** | **0.93** | **0.85** | **81.3** | **3.5** | **0.87** | **0.93** | **0.84** | **83.0** | **3.3** | **0.88** | **0.94** | **0.85** |
| FAST [23] | SIFT [7] | **81.1** | **4.8** | **0.83** | **0.92** | **0.82** | **82.4** | **4.9** | **0.83** | **0.92** | **0.83** | **78.8** | **4.8** | **0.83** | **0.91** | **0.81** | **80.9** | **4.3** | **0.85** | **0.92** | **0.83** |
| FAST [23] | ORB [19] | 64.5 | 25.2 | 0.43 | 0.72 | 0.52 | 68.8 | 27.0 | 0.43 | 0.72 | 0.53 | 70.0 | 23.9 | 0.46 | 0.75 | 0.56 | 70.1 | 26.1 | 0.44 | 0.73 | 0.54 |
| FAST [23] | FREAK [16] | 56.3 | 23.8 | 0.41 | 0.72 | 0.47 | 55.8 | 21.2 | 0.44 | 0.74 | 0.49 | 58.2 | 19.8 | 0.46 | 0.75 | 0.52 | 62.4 | 22.3 | 0.45 | 0.74 | 0.52 |
| FAST [23] | DAISY [17] | 65.5 | 6.0 | 0.76 | 0.88 | 0.71 | 72.7 | 6.5 | 0.77 | 0.89 | 0.75 | 71.9 | 7.4 | 0.74 | 0.88 | 0.73 | 73.0 | 7.0 | 0.75 | 0.88 | 0.74 |
| FAST [23] | BRISK [18] | 60.5 | 21.0 | 0.46 | 0.75 | 0.52 | 64.6 | 29.6 | 0.39 | 0.69 | 0.49 | 66.2 | 21.9 | 0.47 | 0.75 | 0.55 | 71.4 | 28.9 | 0.42 | 0.71 | 0.53 |
| BRISK [18] | BRISK [18] | 100.0 | 99.9 | 0.23 | 0.23 | 0.37 | 100.0 | 99.9 | 0.23 | 0.23 | 0.37 | 100.0 | 99.9 | 0.23 | 0.23 | 0.37 | 100.0 | 99.9 | 0.23 | 0.23 | 0.37 |
| AKAZE [21] | AKAZE [21] | 66.9 | 22.4 | 0.47 | 0.75 | 0.55 | 77.9 | 29.3 | 0.44 | 0.72 | 0.56 | 74.4 | 29.0 | 0.43 | 0.72 | 0.55 | 76.1 | 28.6 | 0.44 | 0.72 | 0.56 |

Table 3. Results of feature point detector and descriptor variants using BoVW with RF classification for firearm detection.

| | | $k_v = 500$ | | | | | $k_v = 1000$ | | | | | $k_v = 1500$ | | | | | $k_v = 2000$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Detector | Descriptor | TP | FP | P | A | F | TP | FP | P | A | F | TP | FP | P | A | F | TP | FP | P | A | F |
| SURF [8] | SURF [8] | **70.7** | **3.5** | **0.86** | **0.91** | **0.77** | **70.9** | **3.5** | **0.86** | **0.91** | **0.78** | **69.9** | **2.8** | **0.88** | **0.91** | **0.78** | **68.4** | **2.5** | **0.89** | **0.91** | **0.77** |
| SIFT [7] | SIFT [7] | 57.9 | 2.6 | 0.87 | 0.88 | 0.69 | 54.8 | 2.1 | 0.88 | 0.88 | 0.68 | 54.0 | 2.1 | 0.88 | 0.88 | 0.67 | 54.4 | 2.3 | 0.88 | 0.88 | 0.67 |
| ORB [19] | ORB [19] | 82.8 | 21.9 | 0.53 | 0.79 | 0.64 | 82.1 | 23.2 | 0.51 | 0.78 | 0.63 | 82.7 | 21.5 | 0.53 | 0.79 | 0.65 | 84.4 | 21.9 | 0.53 | 0.80 | 0.65 |
| KAZE [20] | KAZE [20] | **54.9** | **1.9** | **0.89** | **0.88** | **0.68** | **62.6** | **2.7** | **0.87** | **0.89** | **0.73** | **61.3** | **2.4** | **0.88** | **0.89** | **0.72** | **58.4** | **2.2** | **0.89** | **0.89** | **0.70** |
| FAST [23] | SURF [8] | **71.1** | **3.2** | **0.87** | **0.91** | **0.78** | **69.2** | **2.7** | **0.88** | **0.91** | **0.78** | **72.1** | **3.3** | **0.86** | **0.91** | **0.79** | **71.9** | **2.8** | **0.88** | **0.91** | **0.79** |
| FAST [23] | SIFT [7] | **60.0** | **3.2** | **0.85** | **0.88** | **0.70** | **61.2** | **2.9** | **0.86** | **0.89** | **0.72** | **59.3** | **2.8** | **0.86** | **0.89** | **0.70** | **58.9** | **2.7** | **0.87** | **0.89** | **0.70** |
| FAST [23] | ORB [19] | 90.9 | 26.8 | 0.50 | 0.77 | 0.64 | 90.7 | 28.5 | 0.48 | 0.76 | 0.63 | 91.2 | 28.2 | 0.49 | 0.76 | 0.64 | 91.2 | 28.0 | 0.49 | 0.76 | 0.64 |
| FAST [23] | FREAK [16] | 88.7 | 27.3 | 0.49 | 0.76 | 0.63 | 89.3 | 27.3 | 0.49 | 0.76 | 0.63 | 89.9 | 26.6 | 0.50 | 0.77 | 0.64 | 90.4 | 27.0 | 0.50 | 0.77 | 0.64 |
| FAST [23] | DAISY [17] | 38.5 | 2.9 | 0.80 | 0.84 | 0.52 | 41.9 | 2.5 | 0.83 | 0.85 | 0.56 | 42.3 | 2.6 | 0.83 | 0.85 | 0.56 | 41.4 | 2.6 | 0.82 | 0.85 | 0.55 |
| FAST [23] | BRISK [18] | 90.9 | 27.9 | 0.49 | 0.76 | 0.64 | 91.8 | 29.3 | 0.48 | 0.76 | 0.63 | 91.6 | 28.8 | 0.48 | 0.76 | 0.63 | 91.7 | 29.4 | 0.48 | 0.75 | 0.63 |
| BRISK [18] | BRISK [18] | 99.9 | 99.9 | 0.23 | 0.23 | 0.37 | 99.9 | 99.9 | 0.23 | 0.23 | 0.37 | 99.9 | 99.9 | 0.23 | 0.23 | 0.37 | 99.9 | 99.9 | 0.23 | 0.23 | 0.37 |
| AKAZE [21] | AKAZE [21] | 91.7 | 34.6 | 0.44 | 0.71 | 0.59 | 93.9 | 34.9 | 0.44 | 0.72 | 0.60 | 93.5 | 35.5 | 0.44 | 0.71 | 0.60 | 95.7 | 37.2 | 0.43 | 0.70 | 0.59 |

analysis, is illustrated for a range of realistic and challenging examples. Future work will consider the classification of additional object types, the use of multi-view X-ray imagery and additional BoVW feature pooling representations.

**Acknowledgements:** Partially funded under Innovative Research Call in Explosives and Weapons Detection, 2013 (sponsored under the UK Government CONTEST strategy in partnership with US Dept. Homeland Security, Sci. and Tech. Directorate) and Innovate UK.

## References

[1] M. Baştan, M. R. Yousefi, and T. M. Breuel, "Visual words on baggage x-ray images," in *Computer Analysis of Images and Patterns*, pp. 360–368, 2011.

[2] 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.

[3] M. Baştan, W. Byeon, and T. Breuel, "Object recognition in multi-view dual energy x-ray images," in *Proc. British Machine Vision Conf.*, pp. 130.1–130.11, 2013.

[4] D. Turcsany, A. Mouton, and T. Breckon, "Improving feature-based object recognition for x-ray baggage security screening using primed visual words," in *Proc. Int. Conf. on Industrial Technology*, pp. 1140–1145, 2013.

[5] D. Mery, "X-ray testing by computer vision," in *Proc. Conf. Comp. Vis. and Patt. Rec. Wksp*, pp. 360–367, 2013.

[6] 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.

[7] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comp. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[8] 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.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 27, no. 8, pp. 1265–1278, 2005.

[10] G. Flitton, T. Breckon, and N. Megherbi, "A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery," *Pattern Recognition*, vol. 46, no. 9, pp. 2420–2436, 2013.

[11] A. Mouton, T. Breckon, G. Flitton, and N. Megherbi, "3D object classification in baggage computed tomography imagery using randomised clustering forests," in *Proc. Int. Conf. on Image Processing*, pp. 5202–5206, 2014.

[12] 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.

[13] A. Mouton and T. 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.

[14] S. Akcay, M. Kundegorski, M. Devereux, and T. Breckon, "Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery," in *Proc. Int. Conf. on Image Processing*, pp. 1057 –1061, 2016.

[15] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, "Tackling the x-ray cargo inspection challenge using machine learning," in *SPIE Defense + Security*, p. pp. 98470N, 2016.

[16] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," in *Proc. Conf. Comp. Vis. and Patt. Rec.*, pp. 510–517, 2012.

[17] E. Tola, V. Lepetit, and P. Fua, "DAISY: an efficient dense descriptor applied to wide-baseline stereo.," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 32, no. 5, pp. 815–830, 2010.

[18] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comp. Vis.*, pp. 2548–2555, 2011.

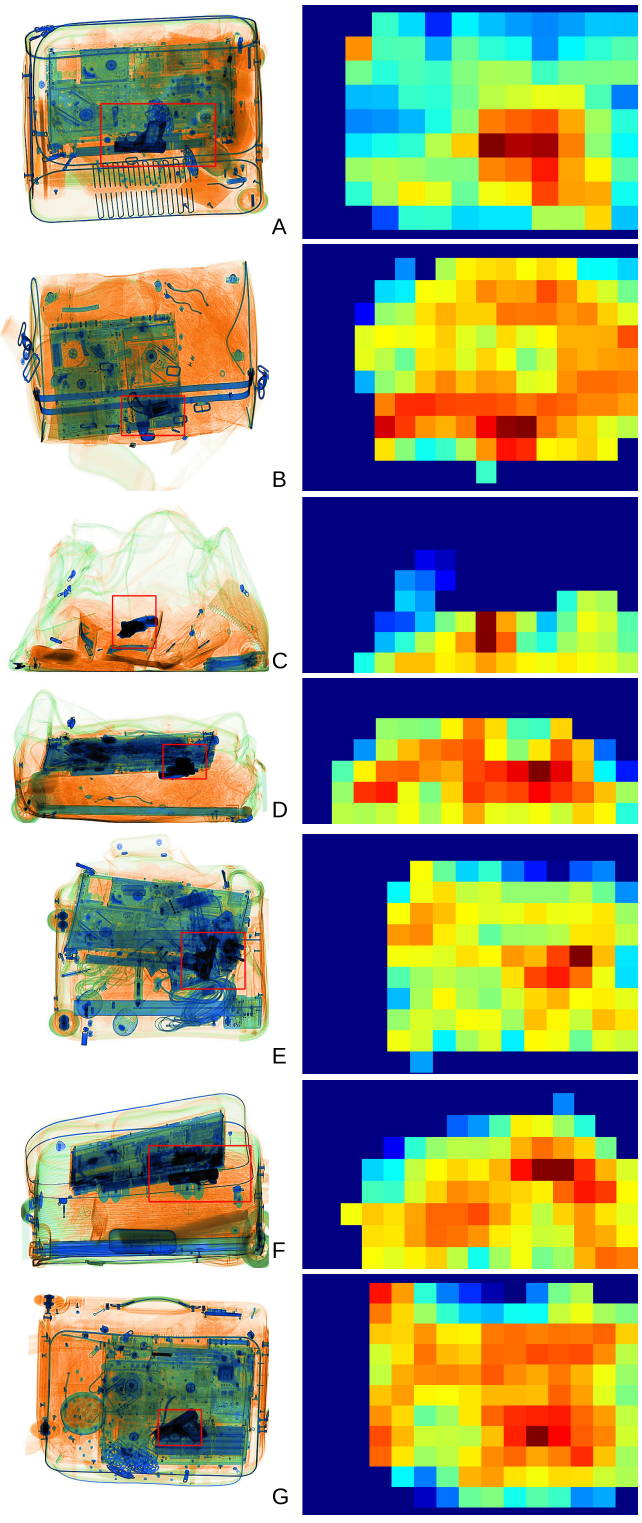


Figure 4. Exemplar detection of whole firearms within cluttered X-ray imagery (FAST-SURF feature detection/description with SVM classification - $\{SVM_{RBF}\}$).


[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *Proc. Int. Conf. Comp. Vis.*, pp. 2564–2571, 2011.

[20] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in *Proc. Euro. Conf. Comp. Vis.*, pp. 214–227, 2012.

[21] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *Proc. British Machine Vision Conf.*, pp. 13.1–13.11, 2013.


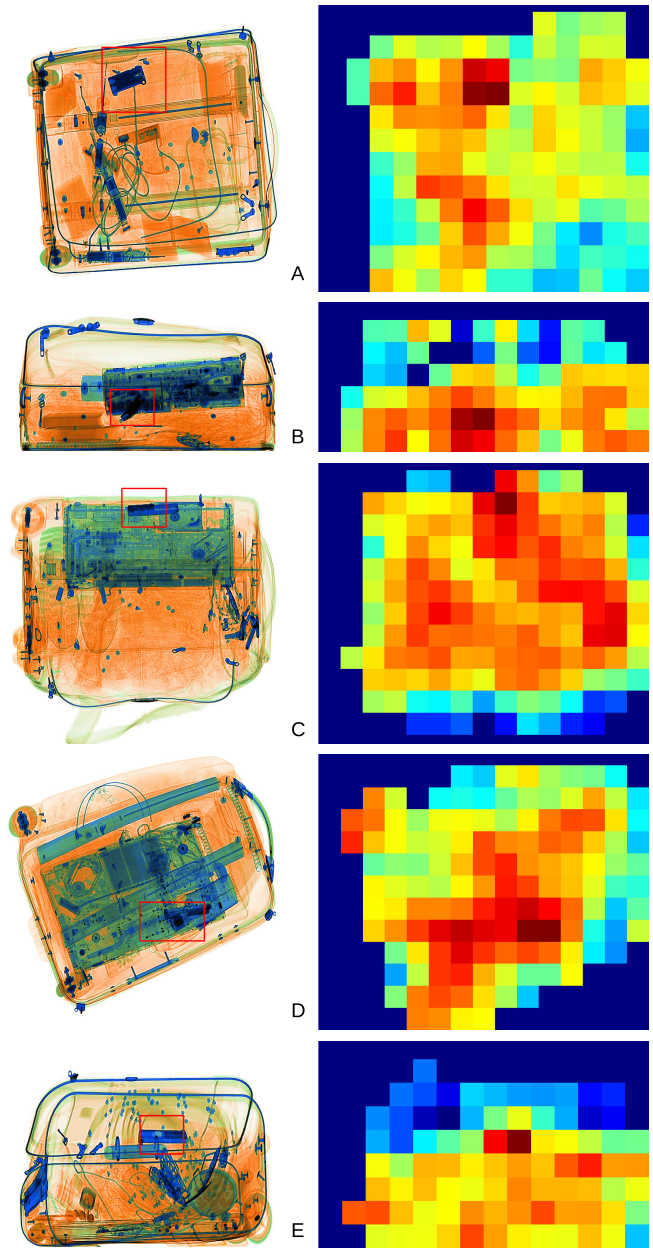


Figure 5. Exemplar detection of firearm sub-components within cluttered X-ray imagery (FAST-SURF feature detection/description with SVM classification - $\{SVM_{RBF}\}$).


[22] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. Int. Conf. Comp. Vis.*, pp. 1470–1477, 2003.

[23] E. Rosten, R. Porter, and T. Drummond, "Faster and better: a machine learning approach to corner detection.," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 32, no. 1, pp. 105–19, 2010.

[24] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "Brief: Computing a local binary descriptor very fast," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 34, no. 7, pp. 1281–1298, 2012.

[25] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 36, 2014.

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

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