Object Recognition in Multi-View Dual Energy X-ray Images

Muhammet Baştan[1]
mubastan@gmail.com
Wonmin Byeon[2]
byeon@rhrk.uni-kl.de
Thomas M. Breuel[2]
tmb@cs.uni-kl.de

[1] Department of Computer Engineering, Turgut Özal University, Ankara, Türkiye (work done while at IUPR)

[2] Image Understanding and Pattern Recognition Group (IUPR), Technical University of Kaiserslautern, Germany

Abstract

Object recognition in X-ray images is an interesting application of machine vision that can help reduce the workload of human operators of X-ray scanners at security checkpoints. In this paper, we first present a comprehensive evaluation of image classification and object detection in X-ray images using standard local features in a BoW framework with (structural) SVMs. Then, we extend the features to utilize the extra information available in dual energy X-ray images. Finally, we propose a multi-view branch-and-bound algorithm for multi-view object detection. Through extensive experiments on three object categories, we show that the classification and detection performance substantially improves with the extended features and multiple views.

1 Introduction

X-ray imaging is a widely used technology at security checkpoints. X-ray images, obtained from X-ray scans, are visually inspected by specially trained human screeners in real-time to detect threat objects. This process is tiring, error prone and also controversial for privacy concerns. Automatic inspection systems using machine vision techniques are not yet commonplace for generic threat detection in X-ray images. Moreover, this problem has not been well explored by machine vision community. This is probably due to lack of publicly available X-ray image datasets, which in turn is due to copyright, security and privacy concerns. In this paper, we address the problem of single view and multi-view object recognition in dual energy X-ray images using local texture/color features and (structural) SVMs.

X-ray security scanners are available for various security applications. Here, we focus on X-ray scanners for check-in and carry-on luggage inspection at airports. Figure 1 shows a simplified model of a 4-view X-ray scanner. As a baggage goes through the scanner tunnel, each X-ray source scans one slice of the baggage at two energy levels (referred to as low energy and high energy), and the attenuated X-rays are collected by a bank of detectors. Then, these slices are combined into a low energy and a high energy grayscale image for each view separately; each slice corresponding to one row in the resultant 2D image. Therefore, the image formation in the direction of belt motion (z) is linear (y on 2D image), while it is

non-linear in the other direction (x on 2D image). The non-linearity as well as the motion of the baggage may cause some distortions in the generated image. At the end of the scanning process, we obtain a total of 8 grayscale energy images for 4 views. These energy images are converted to pseudo-colored RGB images with the help of a look-up table obtained via calibration and displayed to the scanner operators (screeners). The images have a constant width for each view (e.g., 768, 832, 768, 704 pixels for 4 views respectively), but the height is determined by the size of the scanned baggage.

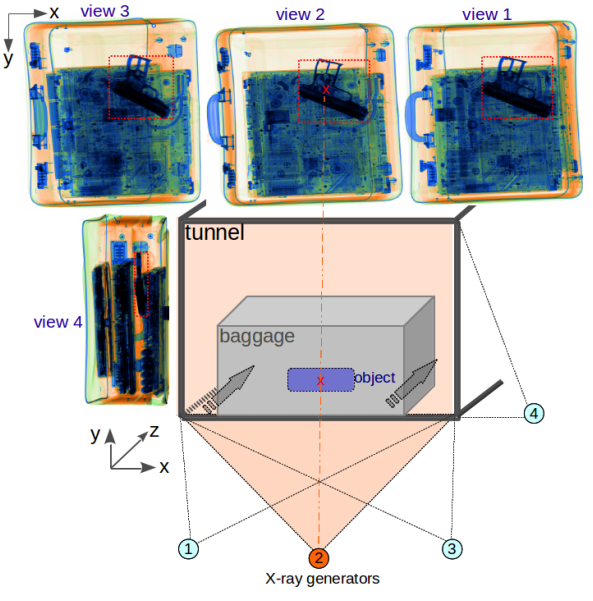

Figure 1: Simplified model of a 4-view X-ray scanner. As the baggage goes through the tunnel (in z direction), 4 X-ray generators scan one slice of the baggage ($x-y$ plane) to generate two energy images for each of the 4 views. The energy images are combined through a look-up table to obtain the pseudo color RGB images for better viewing for the machine operators. The pseudo colors encode the material information (e.g., blue: metals, orange: organic). We make use of two main properties of these X-ray images to boost the object recognition performance: (1) material information, (2) multiple views.

Due to the scanner geometry and imaging process, X-ray images have some peculiarities, which must be considered in the design of automatic machine vision systems for better performance.

Dual energy imaging. The aim of using two energy levels is to obtain both the density and atomic number of the scanned materials. Therefore, the intensity values in energy images and pseudo colors in the color image encode very valuable material information. We show that utilizing this additional information improves object recognition performance significantly.

Multi-view imaging. Multi-view images can help improve the recognition performance. If the object is not visible in one view, it may be visible in another view; this is the motivation for multiple views for human screeners as well. We show that multi-view recognition indeed improves recognition, especially when the single view detections are not very good, as also demonstrated in [7].

Transparency. Transparency may either help or interfere with object recognition; there is no work investigating the effect of transparency on object recognition performance, and this paper also does not address this problem explicitly.

Visual appearance, image content. X-ray images contain much ***less texture*** compared to regular photographic images, with the implication of harder object recognition. X-ray images may get extremely ***cluttered***, with many overlapping, possibly high density objects, making the inspection/recognition impossible even for the human screeners (in which case manual inspection is required). Objects usually undergo in– and out-of-plane ***rotations***, which is another major difficulty for recognition. On the other hand, the problem of change of object ***scale*** and ***illumination*** in regular images is not a big issue in X-ray images, since the machine geometry and energy levels are fixed.

Related Work. The literature on object recognition in X-ray images is very limited. Early works focused more on single view classification and detection/matching [2, 12, 17, 18, 19, 21]. Only recently, methods utilizing multiple views have been proposed [7, 13]. Franzel *et al*. [7] developed an object detection approach for multi-view X-ray images. They used HOG + linear SVM in a brute-force sliding&rotating windows manner in single views; then, they fused the single view detections in 3D to reinforce the geometrically consistent detections while suppressing the false detections. They showed that the multi-view integration improved performance significantly compared to single-view detection. In [6], various 3D feature based approaches were compared and shown that simpler interest point descriptors (density gradient histogram) outperform the complex ones (RIFT and SIFT).

Contributions of this paper. We have three main contributions in this paper. (1) We first present results of extensive image classification and object detection on single view X-ray images using standard local features and (structural) SVMs. (2) Then, we show that both classification and detection performance can be substantially improved with straightforward adaptation and extension of local features (point detectors + descriptors). (3) Finally, we show that the performance can be further improved by utilizing the multiple views and propose a new multi-view branch&bound (B&B) search algorithm for multi-view object detection that uses the single view features directly.

2 Dataset and Feature Extraction

We used an X-ray image dataset from typical 4-view X-ray scanners, like the one in Figure 1. The dataset includes the low energy, high energy and pseudo color images, as well as the exact geometry of the scanners used to record the images. We selected three object classes to perform experiments on: laptops, handguns and glass bottles.

Laptops. Laptops (and similar electronic boards) display a high amount of texture in X-ray images, in contrast to many other objects which are textureless. They are large and have less intra-class variation. These make the laptop recognition easy using standard object recognition algorithms with local features that encode texture information. Our laptop training set has 863 bag scans (131 positive, 731 negative), and test set has 840 bags (119 positive, 721 negative).

Handguns. Handguns contain high density metallic parts as well as non-metallic parts, have less texture compared to laptops, may have large intra-class variations and in– and out-of-plane rotations. These make the handgun recognition difficult. Our handgun dataset contains more than 20 types of handguns in different sizes, some having non-metallic parts, some partly disassembled and some deliberately concealed in the bags. Therefore, it is a

challenging dataset. The training set has 763 bag scans (347 positive, 416 negative), and test set has 921 bags (322 positive, 599 negative).

Glass bottles. There is almost no texture in glass bottles; rotations and intra-class variations complicate the problem even further. Therefore, the shape and material information are two important cues for bottle recognition. Our glass bottle training set has 613 bag scans (88 positive, 525 negative), and test set has 1000 bags (92 positive, 907 negative).

2.1 Feature Extraction

We used local features (interest points, regions, edgels) in a regular bag of visual words (BoW) framework.

Detectors. As local feature detectors, we experimented with both sparse interest points detectors (Harris, Harris-Laplace, Harris-affine, Hessian-Laplace, Hessian-affine) [14, 15, 22] and dense sampling. We used the executables at [14] to detect the sparse interest points, with Hessian and Harris thresholds set to lower values than the defaults to detect more points on textureless X-ray images.

Complementary to sparse interest points, we also experimented with two dense sampling strategies, especially for textureless objects, like bottles: (1) Super pixels (*super*) as an alternative to the grid-based dense sampling, obtained using the SLIC super pixels algorithm [1]. (2) Dense edge sampling (*hedge*), a greedy graph-based clustering of color Canny edges [23] using the color Harris [23] and color edge magnitudes to rank the edge pixels, so that edge segments are clustered based on their curvature. After clustering, the point with the highest score is selected as the representative point for the segment.

Descriptors. To represent the local texture content of the detected points/regions, we used the well-known SIFT descriptor (size 128) and its derivatives (GLOH; CGLOH and CSIFT for color images) [14, 16]. As we noted above, X-ray images contain much less texture than regular images, but the intensity (in energy images) and color (in pseudo color images) encode very valuable material information. We represent this information with the *intensity domain spin image* descriptor (SPIN) [11], a 2D histogram of intensity values in the circular neighborhood of a point. SPIN is rotation invariant. We observed that addition of color feature boosts the classification and detection performance. Therefore, we extend this intensity domain descriptor to multiple channels/images for X-ray images, namely, ESPIN (Energy SPIN) is the concatenation of SPIN descriptors from low and high energy images ($ESPIN = [LSPIN : HSPIN]$), and CSPIN (Color SPIN) is the concatenation of SPIN descriptors from RGB channels of the pseudo color images ($CSPIN = [RSPIN : GSPIN : BSPIN]$). We see that ESPIN works better than SIFT and SPIN alone, and CSPIN is the best. We used the executables at [14] to compute 50-dimensional SPIN descriptors from each image/channel and concatenated them; hence, ESPIN has size 100 and CSPIN has size 150.

3 X-Ray Image Classification

Single view classification. We used the standard BoWs approach: feature detection, feature description, dictionary learning, vector quantization and classification. We extracted the features as described above from low/high energy and pseudo color images, used k-means to construct a dictionary for each feature and for each object class separately ($k = 1000$) and binary SVMs with various kernels (linear, RBF, histogram intersection, chi-square, jensen-

shannon) [4] for classification. To use multiple features, we concatenated the BoW representations of the individual BoW histograms of multiple features.

Multi-view classification. We get multiple images from each X-ray scan. To decide whether an object is present in the scanned bag or not, we can simply classify each view separately and return the best view with the maximum score (*max*), or decide based on the average of the scores from all views (*avg*). Alternatively, we can compute the BoW features from the whole scan, accumulating the contributions from each view and classify the bag as a whole (*bag*).

3.1 Image Classification Results

The experimental results presented here all use the pseudo color X-ray images for interest point detection, since the color images contain more information and texture compared to low/high energy images (they are also more noisy). We performed lots of experiments but could only present the most relevant results due to space limitation. Figure 2 shows the single and multi view X-ray image classification results (binary SVM + histogram intersection kernel). The performance substantially increases with multiple color and texture features, the simpler color features usually performing better than the more complex texture features. Among single descriptors, CSPIN performs the best, indicating the contribution of material information. The classification performance is the highest for laptops and lowest for bottles, as expected (Section 2).

4 Object Detection in X-Ray Images

We adapted the linear structural SVM (S-SVM) with branch-and-bound subwindow search (or efficient subwindow search–ESS) framework of Blaschko and Lampert [3, 10], since it is more efficient and shown to work better than classical sliding windows for object localization in regular images.

S-SVM Training. As in [3], we used the n-slack formulation with margin-rescaling [9, 20], n being the number of training images. We adapted our implementation from the publicly available implementation of SVM-struct at [8] with some modifications to handle multiple ground truth bounding boxes in the training (this might adversely affect the training when there are many training images with multiple bounding boxes, especially when caching is used). We also start with higher resolution (smaller $\varepsilon = 4.0$, instead of the default $\varepsilon = 100.0$) with no loss in performance, and cache the most violating boxes to reduce the training time significantly, from weeks to few hours or days. Finally, instead of the linear loss used in [3], we used quadratic loss based on area overlap (for each ground truth bounding box in a training image separately), since this worked better in our experiments.

Single view branch-and-bound search. The branch-and-bound subwindow search is used to find the subwindow in an image maximizing a quality function. During S-SVM training, it is used to find the most violated constraint in a training image (finding the misdetections with high score using the current model). At test time, using the learned weights in the case of linear S-SVM and with the help of an integral image, it finds the subwindow with the highest score [3]. The authors in [3, 10] parametrize the upright bounding rectangles of objects using by their top, bottom, left and right (t,b,l,r) coordinates, but other parametrizations are also possible; for instance, to impose size constraints on the detected boxes at test time, it is more convenient to use center (x,y), width, height (c_x,c_y,w,h) parametrization. We used (t,b,l,r)

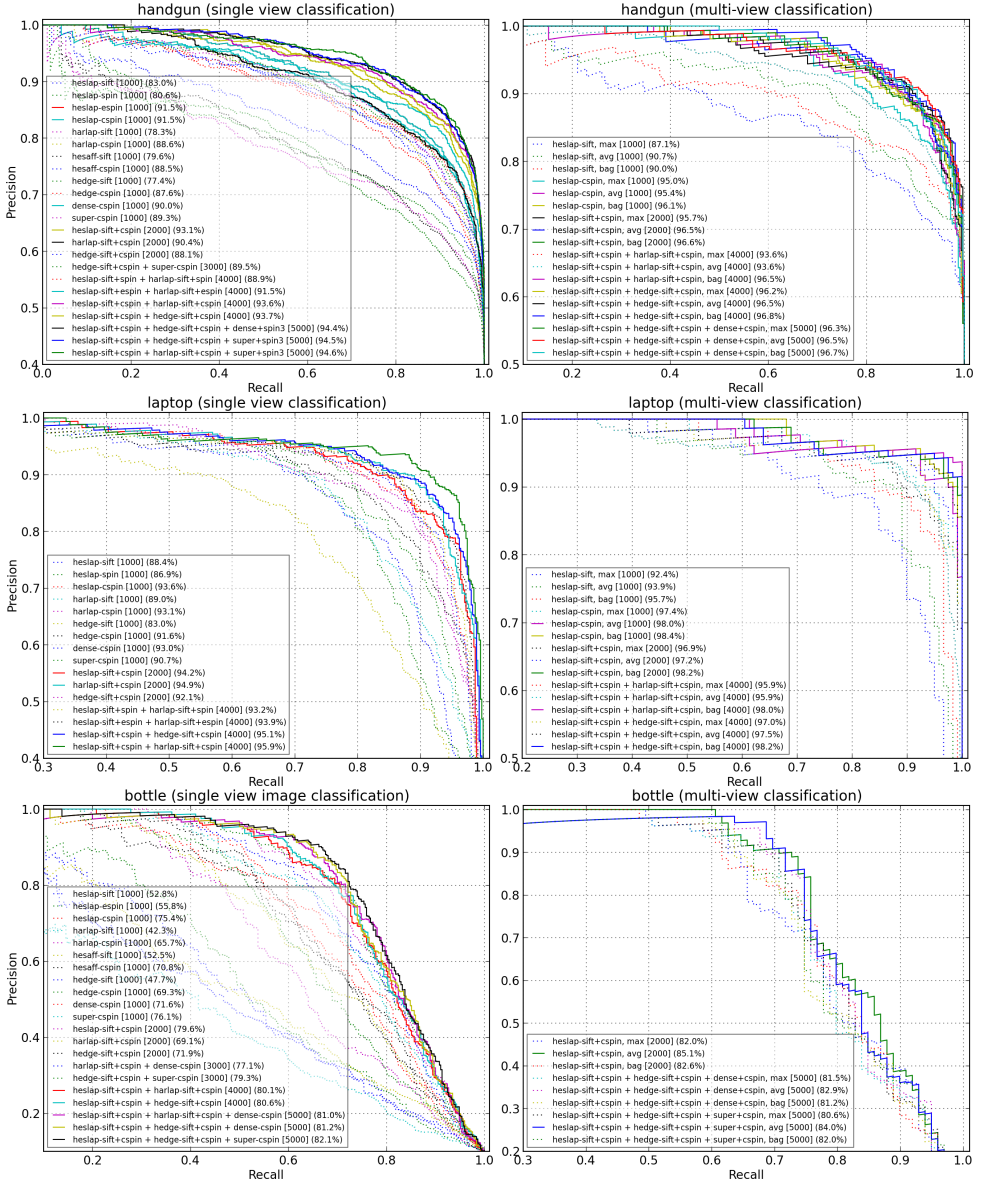

Figure 2: Single view and multi-view X-ray image classification results. The numbers in [] and () are the BoW histogram size and the average precision, respectively. On the right column, *max*, *avg* and *bag* indicate the multi-view classification approach, described in the text.

parametrization during S-SVM training, and (c_x, c_y, w, h) parametrization at test time. This works better, since object sizes in X-ray images do not change much due to fixed machine geometry and hence it is possible to impose size/shape constraints.

4.1 Multi-view branch-and-bound search

In single view detection, each view image of a scan is processed independently, although the object locations are not independent but determined by the machine geometry. As shown in Figure 1, the y coordinates of objects (and hence their heights) are the same (or roughly the same, if some part of the object is not visible due to overlap), while x coordinates are determined by the machine geometry (perspective projection). Intuitively, the object localization accuracy should be improved if the geometry information is utilized. Indeed, multi-view integration has already been shown to improve the detection significantly in [7], by combining the single view detections in 3D. In this paper, instead of combining single view detections, we propose to do the search directly in 3D using the raw features from multiple views. That is, we assume a 3D bounding box around the object in the scanner tunnel (Figure 1), but use features from the projected 2D boxes from the views.

Similar to single view detection, we can parametrize a 3D box with 6 parameters (top, bottom, left, right, front, back). However, we have two problems: (1) when we project axis-aligned bounding boxes onto the views we get larger bounding boxes (see Figure 1), (2) with 6 parameters, the search space becomes huge and the B&B search algorithm does not converge in reasonable time and it also requires too much memory. As a solution, we use an approximation based on the observation that (1) object heights are roughly the same in all views, (2) object widths are very close to each other in the close views (e.g., view 1, view 2, view 3 in Figure 1). Then, we can use the following approximate parametrization center (x, y, z), width, height (c_x, c_y, c_z, w, h) for the 3 close views. Once the multi-view B&B search converges with a rough object location and size (c_x, c_y, c_z, w, h), we can (1) refine the detections on the close views (by allowing e.g., 5-10% change in each parameter), (2) run a single view B&B search on the remaining view image around (c_x, c_y, c_z, h) and estimate w. This way, we are able to obtain a solution with multiple views in reasonable time (seconds).

4.2 Object Detection Results

We used upright bounding box annotations for all object classes in training, and used the same features in detection as in classification (from the pseudo color images). Figure 4 shows the precision-recall graphs for single/multi-view object detection using various features, based on PASCAL's 50% area overlap criterion for a correct detection [5]. From the graphs, we conclude that single texture features perform rather poorly, especially for less textured objects (e.g., bottles). As in classification, the performance improves significantly with the addition of color features (material information). Furthermore, multi-view detection improves the performance considerably especially when the single view detections are not good, as in handguns and bottles. The sample detections in Figure 3 also demonstrates this point; using multiple views improves the localization accuracy, and hence the detection performance.

Comparison. We compared our single view and multi-view detection approach to a recent state-of-the-art approach [7] on handgun detection in multi-view X-ray images. The authors in [7] first annotate their handgun dataset with bounding boxes aligned with the barrel of the gun. Then, they use a HOG-based single view linear SVM detector and fuse the detections

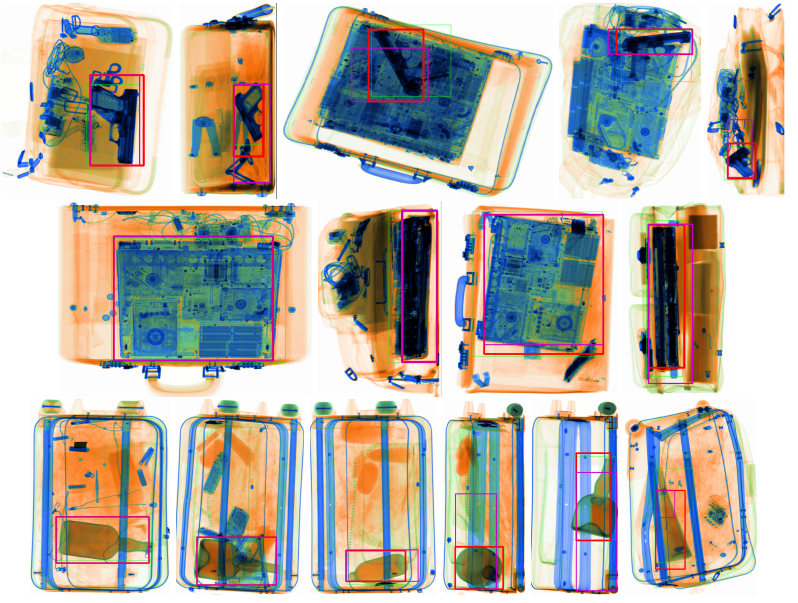

Figure 3: Sample single/multi-view object (handgun, laptop, glass bottle) detections. Green box (thin): ground truth, purple box (medium thickness): single view detection, red box (thick): multi-view detection.

in 3D to reinforce the geometrically consistent detections while suppressing the false detections. Finally, they back-project the detected object locations onto the view images, but lose the concept of rectangular bounding boxes. Therefore, they cannot not use the PASCAL [5] area overlap criterion for evaluation; instead, they define a distance-based criterion: if the distance between the center of ground truth box and detected point is less than $1/3$ of the length of the ground truth box, it is counted as correct detection.

The training set has 514 bag scans (434 positive, 80 negative), and the test set has 255 (216 positive, 39 negative). The precision-recall graph in Figure 5 compares the two approaches. Our approach performs better in terms of both area and distance-based evaluation on single and multi-view detection, although we use simpler, upright bounding box annotations. Moreover, our multi-view algorithm does not lose the concept of rectangular bounding boxes.

5 Acknowledgements

This work was part of the SICURA project supported by the *Bundesministerium für Bildung und Forschung* of Germany with ID FKZ 13N11125 (2010-2013). The X-ray data was provided by Smiths-Heimann (http://www.smithsdetection.com), a manufacturer of X-ray machines and one of the partners in the SICURA project. We thank Pia Dreiseitel, Sebastian Konig, Geert Heilmann, Uwe Siedenburg (Smiths-Heimann), Uwe Schmidt, Thorsten Franzel, Stefan Roth (TU Darmstadt), M. Reza Yousefi, Wanlei Zhao, Ali Bahrainian, Mayce Al-Azawi (IUPR, TU Kaiserslautern) for their help in acquiring and annotating the X-ray images and useful discussions throughout the SICURA project.

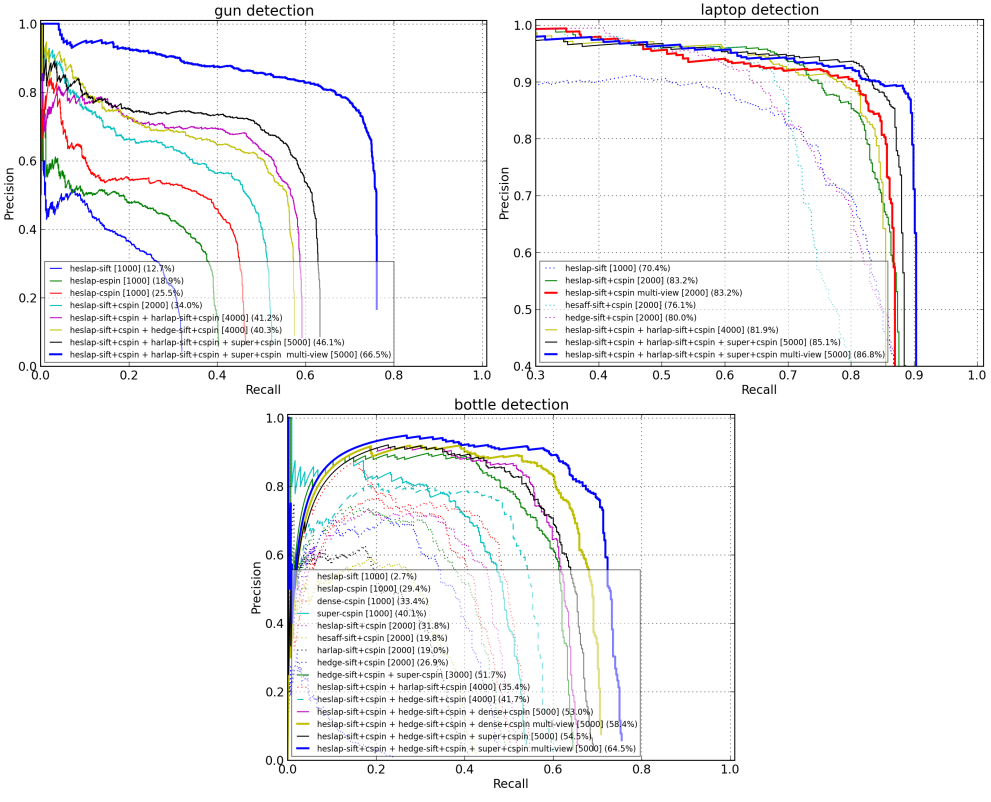

Figure 4: Single/multi-view object detection results. The numbers in [] and () are the BoW histogram size and the average precision, respectively.

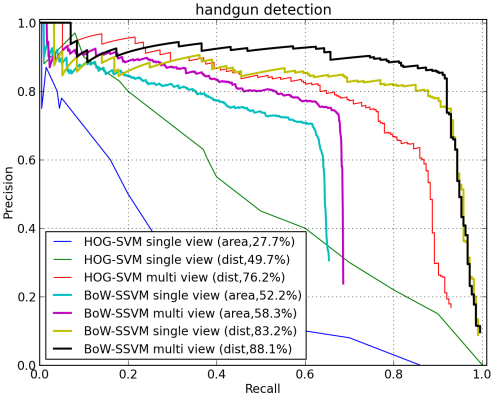

Figure 5: Object detection comparison to [7]. BoW-SSVM is the single/multi-view detectors presented in this paper (with feature 'heslap-sift+cspin + harlap-sift+cspin + super-cspin [5000]'). Inside parenthesis, 'dist' means distance-based evaluation, 'area' means PASCAL's area-based evaluation [5], the numbers are average precision. The single view detection curves are re-generated from [7], while the multi-view detection curves and the distance-based ground truth were kindly provided by the authors [7].

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[2] M. Bastan, M.R. Yousefi, and T.M. Breuel. Visual Words on Baggage X-ray Images. In *CAIP*, 2011.

[3] M.B. Blaschko and C.H. Lampert. Learning to Localize Objects with Structured Output Regression. In *ECCV*, 2008.

[4] C.C. Chang and C.J. Lin. LIBSVM: A library for support vector machines, 2013. URL http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes, 2013. URL http://pascallin.ecs.soton.ac.uk/challenges/VOC.

[6] G. Flitton, T.P. Breckon, and N. Megherbi. A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery. *Pattern Recognition*, 2013.

[7] T. Franzel, U. Schmidt, and S. Roth. Object Detection in Multi-view X-Ray Images. In *DAGM*, 2012.

[8] T. Joachims. SVM-struct: Support Vector Machine for Complex Outputs, 2013. URL http://svmlight.joachims.org/svm_struct.html.

[9] T. Joachims, T. Finley, and Chun-Nam John Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 2009.

[10] C.H. Lampert, M.B. Blaschko, and T. Hofmann. Efficient Subwindow Search: A Branch and Bound Framework for Object Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.

[11] S. Lazebnik, C. Schmid, and J. Ponce. A Sparse Texture Representation Using Local Affine Regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[12] D. Liu and Z. Wang. A united classification system of X-ray image based on fuzzy rule and neural networks. In *International Conference on Intelligent System and Knowledge Engineering*, 2008.

[13] D. Mery. Automated Detection in Complex Objects Using a Tracking Algorithm in Multiple X-ray Views. In *CVPR OTCBVS Workshop*, 2011.

[14] K. Mikolajczyk. Feature Detectors and Descriptors: The State Of The Art and Beyond, 2010. URL http://www.featurespace.org.

[15] K. Mikolajczyk and C. Schmid. Scale & Affine Invariant Interest Point Detectors. *International journal of computer vision*, 2004.

[16] K. Mikolajczyk and C. Schmid. A Performance Evaluation of Local Descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[17] S. Nercessian, K. Panetta, and S. Agaian. Automatic Detection of Potential Threat Objects in X-ray Luggage Scan Images. In *IEEE Conference on Technologies for Homeland Security*, 2008.

[18] C. Oertel and P. Bock. Identification of Objects-of-Interest in X-Ray Images. In *IEEE Applied Imagery and Pattern Recognition Workshop*, 2006.

[19] L. Schmidt-Hackenberg, M.R. Yousefi, and T.M. Breuel. Visual Cortex Inspired Features for Object Detection in X-ray Images. In *ICPR*, 2012.

[20] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Machine Learning for Interdependent and Structured Output Spaces. In *ICML*, 2004.

[21] D. Turcsany, A. Mouton, and T.P. Breckon. Improving Feature-Based Object Recognition for X-ray Baggage Security Screening Using Primed Visual Words. In *International Conference on Industrial Technology*, 2013.

[22] T. Tuytelaars and K. Mikolajczyk. Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision*, 2008.

[23] Joost Van De Weijer, Theo Gevers, and Andrew D Bagdanov. Boosting Color Saliency in Image Feature Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.