Multi-view Object Detection Using Epipolar Constraints within Cluttered X-ray Security Imagery

Brian K. S. Isaac-Medina*, Chris G. Willcocks*, Toby P. Breckon*†
Department of {*Computer Science, †Engineering}, Durham University, Durham, UK

***Abstract*—Automatic detection for threat object items is an increasing emerging area of future application in X-ray security imagery. Although modern X-ray security scanners can provide two or more views, the integration of such object detectors across the views has not been widely explored with rigour. Therefore, we investigate the application of geometric constraints using the epipolar nature of multi-view imagery to improve object detection performance. Furthermore, we assume that images come from uncalibrated views, such that a method to estimate the fundamental matrix using ground truth bounding box centroids from multiple view object labels is proposed. In addition, detections are given a confidence probability based on its similarity with respect to the distribution of the distance to the epipolar line. This probability is used as confidence weights for merging duplicated predictions using non-maximum suppression. Using a standard object detector (YOLOv3), our technique increases the average precision of detection by 2.8% on a dataset composed of firearms, laptops, knives and cameras. These results indicate that the integration of images at different views significantly improves the detection performance of threat items of cluttered X-ray security images.**

***Index Terms*—Multi-view, X-ray security imagery, object detection, epipolar geometry**

I. Introduction

The screening of passenger baggage is an essential task for airport security to avoid threat items entering secure zones. In this regard, the efficiency and aptitude of screening operators is crucial in order to meet the required security standards. Due to the complex and cluttered nature of X-ray security screening imagery, operators must be assessed constantly in order to monitor their performance. Additionally, the ever increasing use of air travel by the public puts increasing pressure on security screening efficiencies. The International Air Transport Association forecasts that the number of air transport passengers could double with up to 8.7 billion passengers globally by 2037 [1]. As a result, the introduction of assistive and automated technologies to aid in the security screening process is a major interest for security [2].

Automatic object detection is a contemporary problem in computer vision, which comprises of the joint localisation and classification of objects of interest within an image with protected objects. Identified objects are usually presented via a bounding box or mask. In this context, deep Convolutional Neural Networks (CNN) have proven to be a reliable technique for object detection [3]–[7]. One detector which has shown a good performance in detection of general objects is R-CNN (regions with CNN features) [8]. The latest iteration of this approach, Faster R-CNN [3], uses a two stage process to predict bounding box detections. First, a region proposal network predicts a set of bounding boxes indicating regions that contain objects. Each of these proposed boxes is given an objectness score, which is the probability of belonging to a set of object categories. In the second stage, these boxes are refined and classified using the architecture described in [9]. Finally, overlapping boxes are merged using non-maximum suppression (NMS) based on their objectness and class probabilities. An alternative competing set of detection architectures can perform joint detection and classification via a single stage, such as SSD [4] or the YOLO family of detectors [10]. YOLOv3 [6] defines a grid over the input image and bounding boxes are parameterized with respect to a set of anchor boxes for each grid cell. As in Faster R-CNN, each box is also assigned objectness and class probability scores with NMS subsequently applied for post-processing.

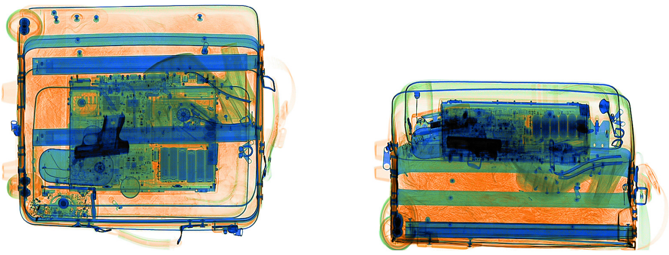

Fig. 1. Exemplar of multi-view X-ray security imagery (bottom/side view).

CNN architectures for object detection have shown to be effective for recognizing threat items in X-ray cabin baggage images [11]–[14]. Different architectures have been tested in X-ray images for threat identification [15], validating their use in this domain. Motivated by the limited availability of X-ray cabin baggage images, transfer learning is used as an initialization step before training [12], [15]. As a result, in this work, we use CNN-based architectures for single view object detection as the basis for extension into multiple view object detection and inter-object resolution.

Contemporary X-ray scanners used for aviation security screening provide two or more views of the baggage content (Figure 1). The geometry of two views of the same scene is related by epipolar geometry [16]. In epipolar geometry, there is a map $x \mapsto l'$ that geometrically relates a point x in one view to a line l' in the other view, known as the epipolar line, where the corresponding point x' lies [16]. Epipolar geometry is exploited in different stereo-vision and multi-view problems, such as visual odometry [17], human pose estimation [18] and 3D reconstruction [19]. Epipolar geometry constraints are encapsulated in the relation $x'^{\intercal}Fx = 0$, where F is known as the fundamental matrix. F can be constructed using the internal parameters of the cameras and its relative position (calibrated cameras), or estimated if a set of point correspondences $\{x_i \leftrightarrow x'_i\}$ is given [16]. When the geometry is unknown (uncalibrated cameras) and point correspondences are not provided, the common methodology is to use feature detectors and descriptors to find matches between the different image views and then proceed to solve for F via least-squares minimization of the geometric inter-image feature projection error [20]. However, the prior work from Kluppel et al. [21] demonstrates that conventional feature detection and matching is not suitable for transmission imagery such as X-ray due to the transparent nature of the object projections which vary with perspective view. Moreover, prior object detection work using multiple view X-ray imagery, with consideration for epipolar constraints, is limited and primarily focuses on 3D bounding box reconstruction [22], where three views are needed [16].

By contrast, this work addresses the use of the epipolar geometry as a constraint to improve the performance of object detection in X-ray security imagery, where perspective view points are uncalibrated and point correspondences are unknown. Our approach leverages the centres of ground truth bounding boxes used for training a contemporary object detection approach [6] as an approximation of point correspondences to estimate the fundamental matrix. Subsequently, the distance of a given bounding box detection from an epipolar line projected from another view is modelled as a random variable with a normal distribution. Finally, the inter-view projection distance of the epipolar line is used to get a multi-view correspondence probability which is jointly used with class and objectness probabilities for subsequent NMS post-processing.

Our key contributions are as follows:

- A novel approach for recovery of the fundamental matrix from uncalibrated views based on the use of ground truth object-level annotations, applied to transmission (X-ray) imagery where conventional feature point matching fails [21].
- Formulation of a multi-view detection approach that cross correlates detections from multiple views by considering the inter-view epipolar constraint as an additional measure of confidence with NMS post-processing.
- Improved benchmark performance for the detection of representative threat objects within x-ray security imagery, based on the correlation of detections across multiple views, outperforming the prior work of [15].

II. Related Work

The first attempts to use the multi-view geometry as a constraint within X-ray security imagery are focused on matching keypoints across the views. One of the earliest works to use multiple views from X-ray imagery is presented by Mery [23]. In this work, objects of interest such as razor blades and pencil tips are segmented using classical feature descriptors and are matched across different views if they lie near a region defined by the epipolar geometry. Fundamental matrix estimation is carried out using point correspondences generated by feature descriptors. Although this method shows a recall of 94.3% and a false positive rate of 5.6%, the test data set is small and samples are not highly cluttered in contrast to the consideration of operational conditions in the X-ray threat object detection work of [24]. A later work from Mery et al. [25] proposes the spatial reconstruction of matched keypoints. Subsequently, these points are clustered and projected back to the 2D domain only if they are large enough. The fundamental matrix is estimated as in [23] and matching keypoints are obtained through a heuristic process. More recent work from the same team on multi-view object detection [26] is a three-step process with deep learning approaches. In the first step, threat objects are detected using similarity of features and spatial distribution. Subsequently, reinforcement learning is used to predict the next view given the object in one source view. Finally, predictions are constrained using the epipolar geometry and the process described in [25]. This method increases the precision of handgun detection from 33% to 84% and the recall from 18% to 66%. Nevertheless, deep CNN object detectors outperform these approaches using single view imagery [11], [15], [24].

In the same context of classical techniques for object detection, Bastan et al. [27] proposed a simple method to search for objects in a spatial domain from 2D raw features. They noticed that in an X-ray machine, bounding boxes of the same object at different views have approximately the same height and the same y coordinate. They take advantage of this constraint, but do not fully exploit the fact that these conditions are an effect of the epipolar geometry (i.e., epipolar lines being almost vertical).

The most recent work on multi-view object detection in X-ray imagery adds a 3D region of interest pooling layer to the Faster R-CNN architecture [22]. This work assumes that the relative position of the viewpoints is known, so scene reconstruction is possible [16]. This method pools deep features of each view into a spatial feature tensor to regress a 3D bounding box. Ground truth 3D bounding boxes are constructed by wrapping the polyhedron formed from the intersection of the rays of projection of 2D bounding boxes. Standard metrics are calculated by re-projecting back the detected 3D bounding box to the 2D domain. They were able to increase the average precision for firearm detection from 85.56% to 92.29%.

In this work we use a deep CNN object detector architecture and filter detections by a constraint imposed to their distance to the epipolar line, such as [26]. However, unlike [22], [25], [26], we give greater weights to bounding boxes with centroids closer to the epipolar line for NMS post-processing. Furthermore, we assume relative position is not known, so a method to estimate the fundamental matrix based on object detection training annotations is proposed.

III. Method

The aim of our proposed approach is to exploit the constraints imposed by the epipolar geometry between the multiple X-ray views in order to improve detection performance. Specifically, we are interested in increasing detection performance whilst reducing false positive detection by correlating across multiple X-ray views and simultaneously improving object localization using the geometric distance of the bounding box to the associated inter-image epipolar lines. In this way, we deal with uncalibrated image viewpoints such that object annotations, available from detector training, are used to estimate the fundamental matrix between these views. The resulting epipolar constraints between views are used to form the basis for subsequent multi-view object detection and filtering.

A. *Fundamental Matrix Estimation*

The fundamental matrix characterizes the epipolar geometry between two views. Given two corresponding points $\{\mathbf{x} \leftrightarrow \mathbf{x}'\}$ in homogeneous coordinates, the fundamental matrix F is a rank-2, 3×3 matrix that satisfies

$$x'^{\intercal} F x = 0\,. \tag{1}$$

F can be obtained using the projection matrices of both cameras and the epipole, which is the point of intersection of the line joining the centres of the cameras (X-ray projection viewpoints) with the image plane [16]. When cameras are uncalibrated or their relative position is not known, F can be exactly calculated if 8 correspondences between the viewpoints are known. If these correspondences are noisy, which is usually the case, then F can be estimated using different error minimization approaches [28]. One of the simplest methods is the normalized 8-point algorithm [29] that solves Equation 1 algorithmically using least squares optimisation with normalised correspondences and a singularity constraint over F. When combined with RANSAC sampling, this technique generally results in a good approximation of F [16].

We consider that two matching points $\{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}$ represent the projection of the same 3D point $\mathbf{X}_i$ in their corresponding image planes. In practical scenarios, the projected point correspondences are noisy and hence we can write the relation between each coordinate of the measured point $\mathbf{x}_i$ to the real projected point $\bar{\mathbf{x}}_i$ as

$$x_i = \bar{x}_i + \Delta x\,, \tag{2}$$

where $\Delta x \sim \mathcal{N}(0, \sigma^2)$ is the error associated with the measurement process of each point coordinate.

Under our conditions, where we have a set of uncalibrated X-ray viewpoints, traditional feature points based matching will fail [21]. The only available information is the ground truth bounding boxes of the threat items used for training the object detection model. Although there are no explicit correspondences, we can instead use the centroids of the bounding boxes as approximations of point correspondences.

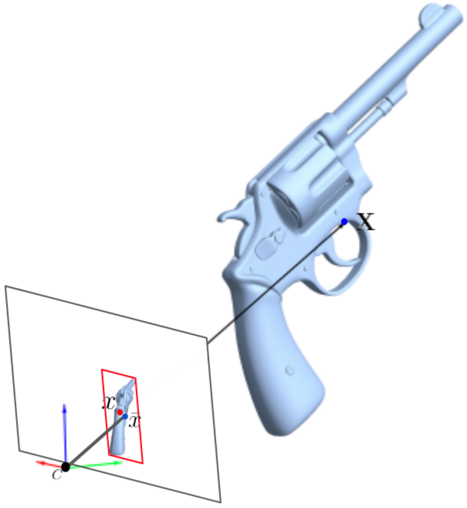

Fig. 2. Comparison of the centre of a bounding box $\mathbf{x}$ with the projection of the real centre of the object $\bar{\mathbf{x}}$ in the plane defined by the camera $\mathbf{C}$.

As seen in Figure 2, there is a difference between the centre of the bounding box with the actual projection of the object centre. This difference is a function of the relative position and orientation of the object with respect to the camera. Hence, the centre of a bounding box $\hat{x}_i$ is modelled as

$$\hat{x}_i = \bar{x}_i + \Delta x + \Psi\,, \tag{3}$$

where Ψ is a function $\Psi : \mathbb{P}^2 \to \mathbb{R}$ that maps the centre of the object in the projective space $\mathbb{P}^2$ of the image to the distance of the centroid of the bounding box. Due to the fact that the position and orientation of the objects are a random event, we can model Ψ as a random variable with a normal distribution $\mathcal{N}(\mu, \sigma'^2)$. Finally, we can write the relation as

$$\hat{x}_i = \bar{x}_i + \Delta \hat{x}\,, \tag{4}$$

with $\Delta \hat{x} \sim \mathcal{N}(\mu, \hat{\sigma}^2)$.

Since $\Delta \hat{x}$ is a function of the object, fundamental matrices are obtained for each object category.

B. *Single View Detection Confidence*

Our method aims to use the epipolar geometry as a constraint for post-processing object detection across multiple views in order to improve global detection performance. Subsequently, approaches to detect are agnostic and hence a standard object detection architecture can be used.

In this work, YOLOv3 [6] is used because it is a fast detector that has shown superior performance in prior work on threat object detection in X-ray images [15]. YOLOv3 defines a detection probability (confidence) which is calculated as

$$P(C = c, O) = P(C = c|O)P(O)\,, \tag{5}$$

where $P(O)$ is the objectness probability, or the probability of the object being an occurrence of one of the object class types considered at training time and $P(C = c|O)$ is the probability of that object being an instance of category c given that it is a valid object. This probability is used as a weight value for NMS post-processing. Within the next section we extend this probabilistic reasoning via consideration of concurrent detections in multiple viewpoints of the same (X-ray) scene that are geometrically correlated by the epipolar constraint between the views.

C. Multiple View Epipolar Detection Confidence

We are now interested in extending the probability associated with each detection to take into consideration concurrent detections from other views. In epipolar geometry, a point position $\mathbf{x_i}$ within one view is projected to a line $\mathbf{l}'$, known as the epipolar line, in the corresponding view using the fundamental matrix F:

$$\mathbf{l}' = F\mathbf{x}_i \,. \tag{6}$$

The distance of a separate point $\hat{\mathbf{x}}'_\mathbf{i}$ within this secondary view to the projected epipolar line is

$$d(\hat{\mathbf{x}}'_i, \mathbf{l}') = \frac{\hat{\mathbf{x}}'^{\intercal}_i \mathbf{l}'}{\sqrt{l'^2_1 + l'^2_2}} = \frac{1}{c}\hat{\mathbf{x}}'^{\intercal}_i \mathbf{l}' \,, \tag{7}$$

where l'^2_1 and l'^2_2 are the first two components of the epipolar line vector and $c = \sqrt{l_1^2 + l_2^2}$. The sign in Equation 7 indicates the half-plane (defined by $\mathbf{l}'$) the point $\hat{\mathbf{x}}'_i$ lies.

Substituting the point coordinates using the relation in Equation 4 into Equation 7 gives:

$$d(\hat{\mathbf{x}}'_i, \mathbf{l}') = \frac{1}{c}\bar{\mathbf{x}}'^{\intercal}_i \mathbf{l}' + \frac{l_1}{c}\Delta\hat{x}_{i1} + \frac{l_2}{c}\Delta\hat{x}_{i2} \,. \tag{8}$$

Assuming that the epipolar line comes from a point with no error[1], the first element of the right side of the previous equation vanishes as the true correspondence point $\bar{\mathbf{x}}_i$ lies in $\mathbf{l}'$. Since the error in both coordinates $\hat{x}_{i1}$ and $\hat{x}_{i2}$ have a normal distribution, we conclude that $d \sim \mathcal{N}(\mu_d, \sigma_d^2)$.

Next, we obtain the probability of object detection bounding box B' in one view belonging to the same object instance as object detection bounding box B in another view based on the distance centroid of B' to the epipolar line defined by the projection of the centroid of B via the corresponding fundamental matrix, F (Equation 6). Therefore, if D is the random variable describing the distance d of the centroid of B' to the epipolar line given by (the projection of the centroid) of B from the corresponding view, the probability p is written as:

$$p(d|B; \mu_d, \sigma_d^2) = P(\mathrm{D} > |d| \cup \mathrm{D} < -|d|; \mu_d, \sigma_d^2) \,, \tag{9}$$

which is the sum of the tails of the probability distribution of D. For a normal distribution, the probability is thus given by

$$p(d|B; \mu_d, \sigma_d^2) = \mathrm{erfc}\left(\frac{d - \mu_d}{\sqrt{2}\sigma_d}\right) \tag{10}$$

where erfc is the complement of the error function. Equation 10 can also be seen as the p-value under the hypothesis that B' is a match of B (under the assumption that the occurrence of threat objects is sparse within the imagery, giving rise to a simplified one-to-one / one-to-few matching problem).

Equation 10 can be used to get an interval of confidence of valid object detection bounding boxes based on their distance to the epipolar line. This method is explored by [26] but with a heuristic decision of the size of the region. Another option is to combine Equations 5 and 10 to get a new extended confidence probability based on the original probabilities from the object detection model and the epipolar constraints between that view and the view containing B. This new confidence probability, which we call multi-view epipolar confidence, is expressed as:

$$\begin{aligned} P(C = c, O, D = d|B) &= P(C = c, O)P(D = d|B) \\ &= P(C = c|O)P(O)p(d|B) \,, \end{aligned} \tag{11}$$

where $P(C = c|O)$ is the class confidence probability given it is a valid object, $P(O)$ is the objectness confidence and $p(d|B)$ is given by Equation 10.

D. Multi-view Filtering

In single view detection, the output of the model is filtered by a its objectness and redundant boxes are removed using NMS [6]. As an extension, we propose a post-processing algorithm that uses the epipolar constraints described in previous sections as an extra step before NMS. We refer to this algorithm as multi-view filtering and the general outline is presented in Figure 3.

First, single view bounding box predictions $\mathrm{B}^m = \{b_i^m\}$ with an objectness confidence probability greater than a threshold value t_o are obtained for a view m. For each b_i^m with category c, we find a set of bounding boxes $\mathrm{B}^n_{m,i} = \{b^n_{m,i,j}\}$ in a different view n with a multi-view epipolar confidence, defined in Equation 11, satisfying

$$P(C = c, O, D = d|B) > rt_c \,, \tag{12}$$

where t_c is a class confidence threshold and r is the minimum p-value of $b^n_{m,i,j}$ as being a correspondence of b_i^m. These boxes are combined using NMS and the resulting bounding box $b^n_{m,i}$ with the greatest multi-view epipolar confidence is considered as the match of b_i^m in the view n. If $\mathrm{B}^n_{m,i}$ is empty for all $n \neq m$, b_i^m is disregarded. Finally, for a dataset with N views, we combine single view predictions and epipolar filtered predictions into a single set of bounding boxes for each view m:

$$\mathbf{B}^m = \mathrm{B}^m \cup \bigcup_{\substack{n=1 \\ n\neq m}}^{N} \bigcup_{b_i^m \in \mathrm{B}^m} b^n_{m,i} \,. \tag{13}$$

Redundancies in $\mathbf{B}^m$ are removed by NMS using their multi-view epipolar confidence as weights for box fusion. The multi-view epipolar confidence of single view predicted boxes is set equal to their class confidences.

As an alternative, we can filter first the bounding boxes within the interval of confidence r and then filter them by their class probabilities (or vice-versa), applying NMS with class probabilities as weights. This technique is similar to the work of [26] and we subsequently explore this process in

[1] We consider only the case where the error comes from the measured point in the same view and not from the epipolar line. The consideration of errors in both sources will be explored in future work.

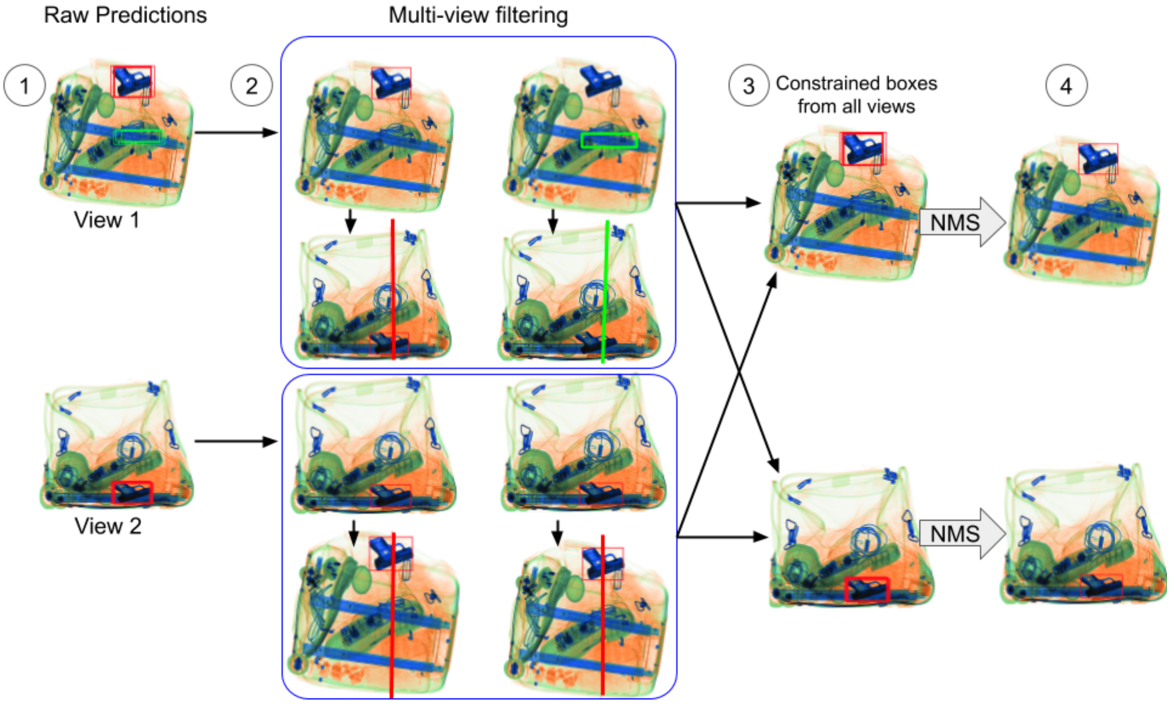

Fig. 3. Multi-view filtering algorithm. This algorithm uses epipolar constraints as a post-processing step before NMS. *1:* Predictions are filtered by objectness, *2:* (a) for each bounding box from step 1 in one view, (b) a set of boxes filtered by their epipolar detection confidence are obtained in the second view. If nothing is found around the epipolar line, then the source bounding box is considered as invalid. *3:* valid bounding boxes from step 2a and filtered boxes from step 2b are combined, *4:* NMS is applied using the epipolar detection confidence.

the ablation studies, showing that the our proposed algorithm yields a better detection performance.

IV. Experimental Setup

In this section, the details about the dataset and the implementation of the object detection model are described.

A. Dataset

Our dataset consists of conventional false-coloured X-ray security imagery from a Smith Detection dual energy scanner with four views (three below and one at a side). We refer to samples as the set of all views of the baggage. A total of 2,528 baggage (10,112 images) were scanned and four object categories were identified. In total, there are samples of 1,090 firearms, 594 laptops, 1,184 knives and 166 cameras. A split of 80% of the samples was used for model training and fundamental matrix estimation. These objects were manually annotated with bounding boxes across all views and a local index was assigned to identify the same object instance across all views. The dataset includes images with only one object and more challenging samples with two or more objects.

B. Evaluation Criteria

We evaluate the performance of our method using MS-COCO detection metrics [30]. The object detection task is evaluated by the number of objects that are identified. A prediction is considered a true positive if the area of intersection over union (IoU) of the ground truth B_{gt} and the prediction B_p is greater than some value. The IoU is given by

$$IoU(B_{gt}, B_p) = \frac{Area(B_{gt} \cap B_p)}{Area(B_{gt} \cup B_p)}.$$

MS-COCO metrics are based on precision and recall over all categories. Precision is the proportion of true positives over all predicted positives while recall is the fraction of correct predictions. Precision increases when IoU increases, while recall tends to decrease. As the value of the IoU increases, precision tends to increase while recall decreases. Average precision (AP) and average recall (AR) are obtained by averaging these values with different values of IoU.

C. Object Detection Implementation

The YOLOv3 [6] object detection architecture is used for single view detection using a DarkNet-53 backbone pretrained on the MS-COCO [30] dataset. Input images are square padded with a white background and resized to 544 × 544. The model is trained using Adam optimization [31] with a learning rate of 0.0001, weight decay of 0.0005, batch size of 8 and for 50 epochs. The learning rate is reduced by a factor of 10 after 15 and 30 epochs. Objectness and class confidence probabilities are set to 0.5, while the minimum p-value for epipolar filtering is 0.05. The model was trained using an Nvidia 2080Ti.

V. Results

In this section we review the results of our proposed methods for fundamental matrix estimation and multi-view

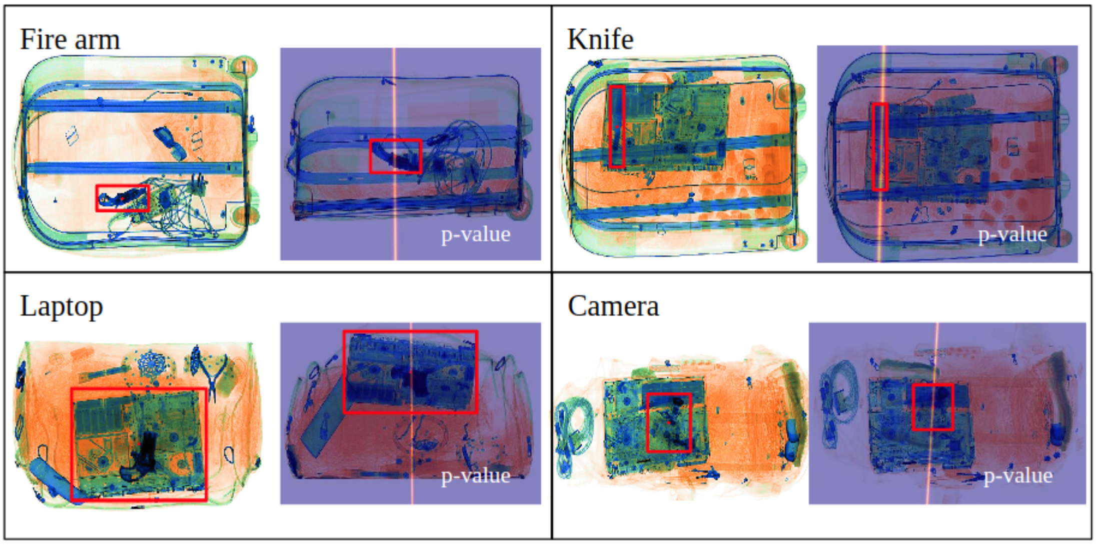

Fig. 4. Results of fundamental matrix estimation per class. The right images for each category show the p-value of the position of candidate bounding box centroids with respect to the epipolar line defined by the left images in another view.

filtering of predictions. Ablation studies are carried out for object detection, modifying some parts of the multi-view filtering algorithm.

A. Fundamental Matrix Estimation

Results for epipolar constraint (line) estimation using the centre of bounding boxes as correspondences are shown in Figure 4. Left images show a ground truth bounding box while right images show the p-value as a function to the distance to the epipolar line defined by the source images, given by the Equation 10. Values for μ_d and σ_d were obtained by fitting the distances of corresponding points and epipolar lines to a normal distribution. It can be seen that some objects such as knives have a wider dispersion. This can be explained by the greater variability of the position of knives in the baggage as compared with bigger objects such as laptops. Also, the error associated with the measurement process (i.e., the bounding box manual annotation) is bigger for smaller objects. These results validate the use of bounding box centres as approximations for inter-view correspondences.

B. Object Detection Performance

Object detection using multi-view filtering is compared against standard single view detection. Table I shows the performance evaluation using COCO metrics for each class as well as metrics for all classes (AP_{100} is not included as our dataset only has up to three objects per image). SV refers to single view YOLOv3 detection and MV to detection processed with multi-view filtering. It is observed that our method outperforms single view detection with an increase of 2.2% of the average precision metric using all categories and IoU values and 2.8% of average precision with a fixed IoU of 0.5. Moreover, multi-view filtering increases marginally all average recall metrics. The improvement of the precision metrics is associated with the elimination of false positives that do not fulfill the epipolar constraints. An example of this elimination is shown in Figure 5a, where an incorrectly identified knife is eliminated after multi-view filtering. Also, sometimes the object detection model has problems detecting overlapping objects from a complex scene. Figure 5b shows a scanned baggage with three overlapping objects where the network identified several objects instead of a single camera. It also predicts two firearms overlapping the a laptop in the lateral view when in the bottom view only one is present. In this case, the multi-view filtering algorithm successfully removes all incorrect items that were identified near the camera and removes the false firearm overlapping the laptop. However, in a more challenging scenario, such as the scanning from Figure 5c, some correct detections are removed because they were not found in the other views.

A further analysis of the results shows that our method is not optimized to long objects such as laptops or large knives. This is indicative that the assumptions made for Equation 4 about the nature of Ψ may be invalid and a more detailed analysis is necessary. The modelling of the relation between the projected object centre and the centre of a bounding box in an image plane is left for future work.

C. Ablation Studies

In this section we want to assess experimentally how our methodological choices perform compared against some simplifications. We focus in two main parts of our method: modelling the distance between a bounding box and an epipolar line as a normal distribution with mean μ_d (not necessarily equal to zero) and the use of the multi-view epipolar confidence in (11) for NMS.

First we validate the choice of modelling our (signed) distance of the bounding box to the epipolar line with a biased estimator, i.e., $d \sim \mathcal{N}(\mu_d, \sigma_d^2)$. To do so, we run a test assuming that the distance follows a normal distribution with 0 mean. As can be seen in the third row of Table II, using a unbiased estimation of the distance, multi-view filtering

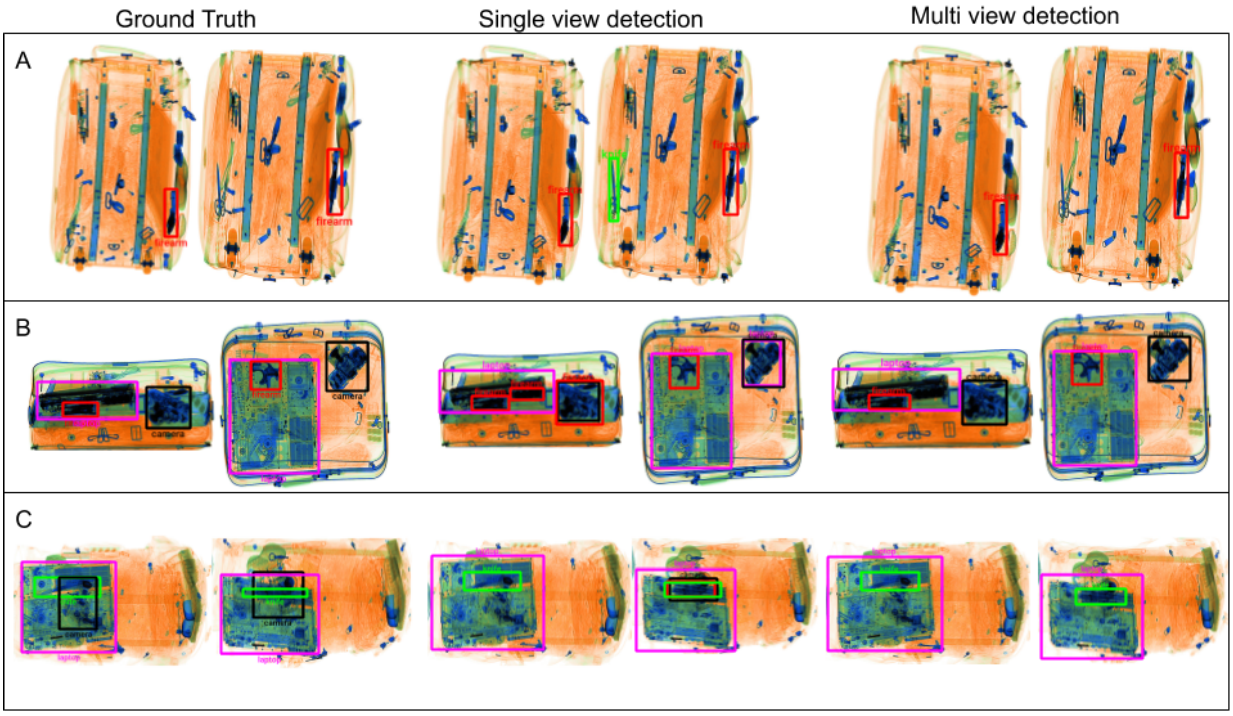

Fig. 5. Comparison between ground truth, single view and multi-view detection. Object categories are labelled by colour: red for firearms, green for knives, fuchsia for laptops and black for cameras. (A) Elimination of false positives. (B) Elimination of more complex false positives and correction of class. (C) Missing of a previous identified class.

TABLE I
MULTI-VIEW OBJECT DETECTION RESULTS

Category	Method	AP	AP$_{0.5}$	AP$_{0.75}$	AP$_S$	AP$_M$	AP$_L$	AR$_1$	AR$_{10}$	AR$_S$	AR$_M$	AR$_L$
Firearm	SV	0.670	0.983	0.816	-	0.681	0.630	0.743	0.747	-	0.744	**0.776**
	MV	**0.691**	**0.988**	**0.848**	-	**0.702**	**0.679**	**0.746**	**0.749**	-	**0.747**	0.775
Laptop	SV	**0.705**	0.972	**0.886**	-	-	**0.705**	**0.770**	**0.772**	-	-	**0.772**
	MV	0.697	**0.973**	0.872	-	-	0.697	0.764	0.766	-	-	0.766
Knife	SV	0.320	0.726	0.236	0.083	0.349	**0.175**	0.440	0.447	0.112	0.464	0.263
	MV	**0.382**	**0.800**	**0.322**	**0.125**	**0.412**	0.138	**0.455**	**0.463**	**0.154**	**0.478**	**0.287**
Camera	SV	0.530	0.848	0.621	-	**0.700**	0.530	**0.605**	**0.605**	-	**0.700**	**0.605**
	MV	**0.546**	**0.881**	**0.633**	-	**0.700**	**0.546**	0.603	0.603	-	**0.700**	0.602
All	SV	0.557	0.882	0.640	0.083	0.577	0.510	0.640	0.643	0.112	0.636	0.604
	MV	**0.579**	**0.910**	**0.669**	**0.125**	**0.605**	**0.515**	**0.642**	**0.645**	**0.154**	**0.641**	**0.608**

performs worse in all metrics. The reason behind using a biased estimator for the distance is that the mean μ_d serves as a correction of the unknown distance Ψ of the centres of the bounding box with the actual projection of the object centre in the image plane. Subsequently, if we do not induce this bias, the multi-view filtering algorithm looks for matches in a region further away from the actual match.

Secondly, we compare our filtering algorithm by multi-view epipolar confidence with class confidence filtering bounded by a region of confidence. In this case, instead of using the relation in Equation 12 for filtering and performing NMS, we test a model that only looks in the second view for bounding boxes in an interval and then uses the class confidence as weights for NMS, as in single view detection. This method is partially explored by [26], but choosing the interval of confidence heuristically. The results are shown in the first and second rows of Table II, with both biased and unbiased estimators of the distance. Again, this method performs poorly against the superior performance offered by our approach. This is explained noting that the use of the multi-view epipolar confidence defined in Equation 11 gives a greater weight to bounding box detections that are closer to the epipolar line, resulting in higher quality bounding boxes after NMS.

VI. CONCLUSION

In this work we have developed a new multi-view detection approach using epipolar constraints as an additional confidence probability for NMS. The distance of bounding box centroids from corresponding epipolar lines is modelled as a random variable with a normal distribution and non-zero mean. The p-value of the distance with respect to that distribution is used as a new confidence probability for NMS post-processing. Furthermore, a novel method is proposed that estimates the fundamental matrix by making use of ground truth object annotations available from object detector model training.

We show that using bounding box centroids as point correspondences across views allows for high-quality

TABLE II
ABLATION STUDIES

Method	$\mathbf{AP}$	$\mathbf{AP_{0.5}}$	$\mathbf{AP_{0.75}}$	$\mathbf{AP_S}$	$\mathbf{AP_M}$	$\mathbf{AP_L}$	$\mathbf{AR_1}$	$\mathbf{AR_{10}}$	$\mathbf{AR_S}$	$\mathbf{AR_M}$	$\mathbf{AR_L}$
MV class confidence, $d \sim \mathcal{N}(0, \sigma_d^2)$	0.577	0.904	**0.671**	0.085	0.601	0.514	0.641	0.643	0.107	0.636	0.604
MV class confidence, $d \sim \mathcal{N}(\mu_d, \sigma_d^2)$	0.577	0.904	0.669	0.094	0.601	**0.515**	0.641	0.643	0.110	0.637	0.607
MV epipolar confidence, $d \sim \mathcal{N}(0, \sigma_d^2)$	0.576	0.909	0.666	0.099	0.569	0.512	0.640	0.644	0.132	0.606	0.606
MV epipolar confidence, $d \sim \mathcal{N}(\mu_d, \sigma_d^2)$	**0.579**	**0.910**	0.669	**0.125**	**0.605**	**0.515**	**0.642**	**0.645**	**0.154**	**0.641**	**0.608**

estimation of the fundamental matrix. Our approach increases the average precision of the MS-COCO metric by 2.2% and by 2.8% when using a fixed intersection over union of 0.5. Additionally, we find that our proposed method outperforms the approach of simply constraining the bounding boxes to lie in a region around the epipolar line. These results show that the use of epipolar constraints for multi-view object detection is a key contribution for decreasing false positives and improving detection performance in the context of cluttered X-ray security imagery.

Future work will investigate the use of epipolar constraints on different contexts and more complex models for the estimation of the fundamental matrix using object annotations.

VII. ACKNOWLEDGMENTS

This work is partially supported by the Mexican Council of Science and Technology (CONACyT).

REFERENCES

[1] International Air Transport Association, "IATA forecast predicts 8.2 billion air travelers in 2037," https://www.iata.org/pressroom/pr/Pages/2018-10-24-02.aspx, Oct 2018.

[2] O. E. Wetter, "Imaging in airport security: Past, present, future, and the link to forensic and clinical radiology," *Journal of Forensic Radiology and Imaging*, vol. 1, no. 4, pp. 152 – 160, 2013.

[3] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science*, p. 21–37, 2016.

[5] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[6] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[7] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, pp. 1–21, 01 2019.

[8] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524

[9] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015. [Online]. Available: http://arxiv.org/abs/1504.08083

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[11] S. Akcay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging," arXiv: 2001.01293, 2020.

[12] 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*. IEEE, September 2016, pp. 1057 –1061.

[13] S. Akcay and T. Breckon, "An evaluation of region based object detection strategies within x-ray baggage security imagery," in *Proc. Int. Conf. on Image Processing*. IEEE, September 2017, pp. 1337–1341.

[14] K. J. Liang, J. B. Sigman, G. P. Spell, D. Strellis, W. Chang, F. Liu, T. Mehta, and L. Carin, "Toward automatic threat recognition for airport x-ray baggage screening with deep convolutional object detection," 2019.

[15] S. Akcay, M. Kundegorski, C. Willcocks, and T. Breckon, "On using deep convolutional neural network architectures for automated object detection and classification within x-ray baggage security imagery," *IEEE Transactions on Information Forensics & Security*, vol. 13, no. 9, pp. 2203–2215, September 2018.

[16] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, ISBN: 0521540518, 2004.

[17] V. Prasad, D. Das, and B. Bhowmick, "Epipolar geometry based learning of multi-view depth and ego-motion from monocular sequences," in *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, ser. ICVGIP 2018. New York, NY, USA: Association for Computing Machinery, 2018.

[18] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu, "Epipolar transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7779–7788.

[19] Z. Fang, X. Guo, X. Zhu, R. Zhou, and H. Ma, "Sefm: A sequential feature point matching algorithm for object 3d reconstruction," *ArXiv*, vol. abs/1812.02925, 2018.

[20] L. Wang, Z. Liu, and Z. Zhang, "Efficient image features selection and weighting for fundamental matrix estimation," *IET Computer Vision*, vol. 10, no. 1, pp. 67–78, 2016.

[21] M. Klüppel, J. Wang, D. Bernecker, P. Fischer, and J. Hornegger, *On Feature Tracking in X-Ray Images*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 132–137.

[22] J.-M. O. Steitz, F. Saeedan, and S. Roth, "Multi-view x-ray r-cnn," in *Pattern Recognition*, T. Brox, A. Bruhn, and M. Fritz, Eds. Cham: Springer International Publishing, 2019, pp. 153–168.

[23] D. Mery, "Automated detection in complex objects using a tracking algorithm in multiple x-ray views," in *Computer Vision and Patter Recognition 2011 Workshops*, 2011, pp. 41–48.

[24] N. Bhowmik, W. Q., Y. Gaus, M. Szarek, and T. Breckon, "The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composite x-ray imagery," in *Proc. British Machine Vision Conference Workshops*. BMVA, September 2019, pp. 1–8. [Online]. Available: http://community.dur.ac.uk/toby.breckon/publications/papers/bhowmik19synthetic.pdf

[25] D. Mery, V. Riffo, I. Zuccar, and C. Pieringer, "Automated x-ray object recognition using an efficient search algorithm in multiple views," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 368–374.

[26] V. Riffo, S. Flores, and D. Mery, "Threat objects detection in x-ray images using an active vision approach," *Journal of Nondestructive Evaluation*, vol. 36, 09 2017.

[27] M. Bastan, W. Byeon, and T. M. Breuel, "Object recognition in multi-view dual energy x-ray images," in *British Machine Vision Conference*, vol. 1, no. 2, 2013, p. 11.

[28] M. E. Fathy, A. S. Hussein, and M. F. Tolba, "Fundamental matrix estimation: A study of error criteria," *Pattern Recognition Letters*, vol. 32, no. 2, p. 383–391, Jan 2011.

[29] R. I. Hartley, "In defense of the eight-point algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 580–593, 1997.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Equationft COCO: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.