# Threat Objects Detection in X-ray Images Using an Active Vision Approach

**Vladimir Riffo**[1] · **Sebastian Flores**[1] · **Domingo Mery**[2]

Received: 18 January 2017 / Accepted: 28 April 2017

**Abstract** X-ray testing for baggage inspection has been increasingly used at airports, reducing the risk of terrorist crimes and attacks. Nevertheless, this task is still being carried out by human inspectors and with limited technological support. The technology that is being used is not always effective, as it depends mainly on the position of the object of interest, occlusion, and the accumulated experience of the inspector. Due to this problem, we have developed an approach that inspects X-ray images using active vision in order to automatically detect objects that represent a threat. Our method includes three steps: detection of potential threat objects in single views based on the similarity of features and spatial distribution; estimation of the best-next-view using Q-learning; and elimination of false alarms based on multiple view constraints. We tested our algorithm on X-ray images that included handguns and razor blades. In the detection of handguns we registered good results for recall and precision (*Re* = 67%, *Pr* = 83%) along with a high performance in the detection of razor blades (*Re* = 82%, *Pr* = 100%) taking into consideration 360 inspections in each case. Our results indicate that non-destructive inspection actively using X-ray images, leads to more effective object detection in complex environments, and helps to offset certain levels of occlusion and the internal disorder of baggage.

**Keywords** X-ray testing · Threat objects detection · Active vision · X-ray images · Computer vision

## 1 Introduction

Over the last few years, aviation security screening using X-ray scanners has become a very important issue in airports and safety checkpoints. The inspection process, however, is complex as dangerous items are very difficult to detect when placed in closely packed bags, are superimposed by other objects and/or are rotated showing an indistinguishable profile. In baggage screening, where human security plays an important role and inspection complexity is very high, human inspectors are still employed. However, human inspection of threat objects is:

(i) *Demanding and stressful:* During peak hours at airports, inspectors only have a few seconds to decide if a piece of luggage contains or not an element that could be a threat.
(ii) *Boring and tedious:* Very few pieces of baggage actually contain threatening articles. The job requires a lot of focus to identify a wide variety of objects and their categories, forms and substances (metallic, organic and inorganic substances).
(iii) *Difficult:* Human inspectors have to undergo a training program and receive minimal technological support.
(iv) *Uncertain:* Because each operator must examine many and varied bags, packages and luggage, the probability of human error rises considerably over an extended period of time. Reported detection performance is only in the range of 80–90%.

Alternatively, automated X-ray object recognition can offer the advantages of objectivity and reproducibility for every test. For this reason, certain digital imaging and computer vision techniques have been developed. However, even though various scientific teams are exploring numer-

✉ Vladimir Riffo
vladimir.riffo@uda.cl

[1] Departamento de Ingeniería Informática y Ciencias de la Computación, Universidad de Atacama, Copiapó, Chile

[2] Department of Computer Science, Pontificia Universidad Católica de Chile, Santiago, Chile

Published online: 31 May 2017

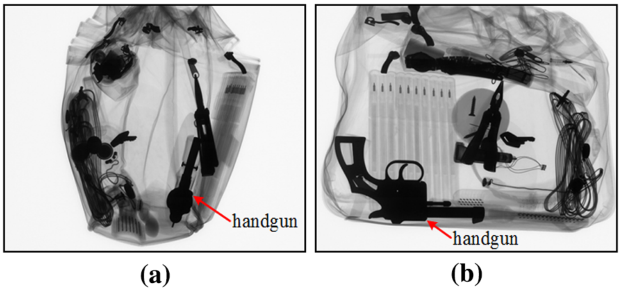


**Fig. 1** Two views of an object inspected with X-rays, **a** handgun in a bad pose, and **b** handgun in a good pose

ous research directions, adopting very different principles, and developing a wide range of algorithms for very different applications, automated X-ray object recognition remains an open question due to the large variability of the appearance and shape of test objects, both between and within categories (e.g., within the category *guns* and *knives* it is possible to identify many different objects). Furthermore, there exists a broad variability within an object sample depending on its points of view (e.g., the overhead view and frontal view of a *gun* are very different). In addition, the appearance of a test object can differ due to self-occlusion, noise and conditions of acquisition. Furthermore, certain threat objects are not easily recognizable after only one view (due to their position as shown in Fig. 1a).

In this research project, we propose a method for automated inspection employing multiple views of X-ray images using a strategy called *active vision*. The key idea of the method is to identify a "good view" to ensure detection as shown in Fig. 1b.

Active vision for X-ray baggage inspection was originally proposed in our preliminary work with promising results in the detection of razor blades placed inside different container objects [1]. The approach attempted to locate a *good view* of the inspection object, i.e., an image in which a target object should be viewed from a *good pose* that ensures its detection. The good poses of the target object correspond to those in which the acquired view should have a high probability of detection as shown in Fig. 1. Thus, the strategy attempts to rotate and/or translate the inspection object from an initial to a new position in which the detection probability of the target object should be higher. In this new research work, we have significantly improved the original active vision approach by:

(i) *Increasing the robustness of single-view detection:* We have employed an Adapted Implicit Shape Model in our work (AISM) [2], which is an adaption of the Implicit Shape Model (ISM) [3]. AISM has been shown to obtain a good performance in baggage inspection as it considers objects as a set of independent parts, but that are connected logically through a star structure that allows us to detect different categories of objects;

(ii) *Improving the prediction of the next view:* In order to predict the next view of the active vision approach, we now use a reinforcement learning algorithm (*Q-Learning*);

(iii) *Including an additional validation step:* We have eliminated false alarms by using an algorithm of geometric correspondences between images, as a hypothetical detection in a single view would be considered real if a correspondence exists in another image.

Finally, in our research work the evaluation of the approach has been undertaken in more realistic environments, e.g., bags with a high degree of clutter using GDXray database [4]. During this process, active inspection has been performed using X-ray images acquired under basic conditions: shades of gray (no pseudocolor), with a single viewpoint, acquired images with only one energy level and with no algorithm for image processing.

We have implemented and applied a method for the automatic and active detection of threat objects placed inside luggage bags, handbags, etc., considering the following as such objects: razor blades and handguns. We have highlighted the robustness of this approach in the active detection of a symmetrical and regular object (razor blades), and acceptable results for objects of irregular form (handguns). These are promising results and establish a generic approach for X-ray image detection of objects, which could be a useful tool to help human inspectors in airports or customs inspection points. We also believe that—with an ad-hoc training dataset—, our method could be useful in detecting other kinds of objects in X-ray images.

## 2 State of the Art

In general, X-ray inspection can be carried out by human inspectors or by automatic systems. Although human inspectors can do the task better than machines, they are slower and tire quickly. Additionally, human inspectors are not always consistent and effective when evaluating objects, as inspection tasks are monotonous and tedious, even for experts. Moreover, human experts are difficult to find or keep within the industry; they require training and their learning process can take time. According to the literature, human inspection of industrial processes has an efficiency level of maximum 80% [5], and in other publications associated with X-ray inspections in airports, the efficiency level does not rise above 90% [6–8].

After the 9/11 terrorist attacks on the United States, airports and customs halls have intensified restrictions and stepped up security. In addition, safety checkpoints using

X-rays have been placed at international or regional borders to avoid the introduction of diseases and pests that can harm local agriculture (i.e., to detect plant and animal products). Thus, inspection through X-ray screening has become a process of high importance. Nonetheless, inspection of handbags, suitcases and other luggage in general is a complex task, as many objects deemed a threat are difficult to detect, especially when packages are placed too close to each other, leading to the occlusion or distortion by other objects, and creating an unrecognizable viewpoint [9–12].

In the course of this work, it has been particularly interesting to review the advances in baggage screening that have taken place over the course of the current decade.[1] These can be summarized as follows: some approaches attempt to recognize objects using a single view of mono-energy X-ray images (e.g., the adapted implicit shape model based on visual codebooks [2], adaptive sparse representations [14] and deep learning [15]) and dual-energy X-ray images (e.g., Gabor texture features [16], bag of words based [17,18], pseudocolor, texture, edge and shape features [19,20] and deep learning [21]). Furthermore, complex approaches that deal with multiple X-ray images have also been developed. For the recognition of regular objects from mono-energy images, methods such as data association [22,23] and active vision [1], where a second-best view is estimated, have been explored. In the case of dual-energy imaging, visual vocabularies and SVM classifiers have been used, as shown in [24]. Progress has also been made in the area of computed tomography. For example, in order to improve the quality of CT images, metal artifact reduction and de-noising [25] techniques have been suggested. Many methods based on 3D features for 3D object recognition have been developed (see, for example, RIFT and SIFT descriptors [26], 3D Visual Cortex Modeling, 3D Zernike descriptors and histograms of the shape index [27]). Contributions have also been made using known recognition techniques (see, for example, bag of words [28] and random forest [29]).

But as we know, existing technology is far from perfect. Today there is no completely automatic method, and manual systems remain vulnerable to human error. The state of the art shows that in this area of research there have been different approaches, depending on the application. Nevertheless, X-ray automatic inspection still has problems that need to be addressed: (i) *loss of generality* this is because approaches developed for one application cannot be used on others, (ii) *poor accuracy in the detection* as the false positive (false alarms) and non detection are compromised, (iii) *limited robustness* due to the fact that pre-requisites for the use of a method are frequently obtained only with simple structures, and (iv) *low adaptability* as it can be very hard to modify the design of an automatic system. We have observed that when inspectors make a baggage radioscopic inspection at airports, they have only one X-ray image (incorporating shades of gray and/or pseudocolor) to take a decision of high importance, i.e., to identify a number of threat objects (metal, organic and inorganic) that could place the lives of people that travel by plane at risk. The literature reports that for these purposes, one image is insufficient [30], given that objects of concern (threat objects), can be totally or partially occluded and/or positioned in such a way that does not allow for their recognition as illustrated in Fig. 1. Multiple views and active vision can be an effective option for examining complex objects where uncertainty can lead to misinterpretation.

## 3 Proposed Method

The general approach allows us to find an image in which a target object can be observed from a ‘good pose’ that assures its detection. For example, the ‘good poses’ for a handgun correspond to the frontal views (the largest visible surface) as illustrated in Fig. 1b. In contrast, in the case of a ‘bad pose’ the object is difficult to identify as shown in Fig. 1a. Our approach follows Fig. 2: the key idea is to rotate the object being examined—in case it is positioned in a ‘bad pose’—from an initial position to a new one, a ‘good pose’, in which the probability of detection of a threat object is increased. Given the acquisition of a primary image, the following three situations may occur:

- *Detection from a ‘good pose’:* If the initial pose corresponds to a good viewpoint and the target object is detected, it will not be necessary to move (rotate) the object to a new position (given by $L$ in Fig. 2), i.e., in this case, the inspection is performed with only one X-ray image, avoiding the need to analyze more images.
- *Detection from a ‘bad pose’:* If a potential target object is detected in a ‘bad pose’, an algorithm will estimate the position $L$ to which the gripping and rotation systems have to move, in order to obtain a ‘good pose’ in the next detection, i.e., the next-best-view. This process is interrupted after $C_{max} = 4$ times, in order to avoid infinity loops.
- *No object detection:* If no target object is identified during the detection stage, then the testing object is arbitrarily rotated to a new position $L$, that is different from the first, and repeating the detection stage. The previous situation may be repeated $D_{max} = 3$ times, so as to guarantee the inspection of all relevant viewpoints.

The proposed approach analyzes X-ray images to detect a target object in a good pose. Our active vision method

[1] Other contributions have been made towards computer vision for X-ray testing in applications such as the inspection of castings, welds, food and cargos [13].

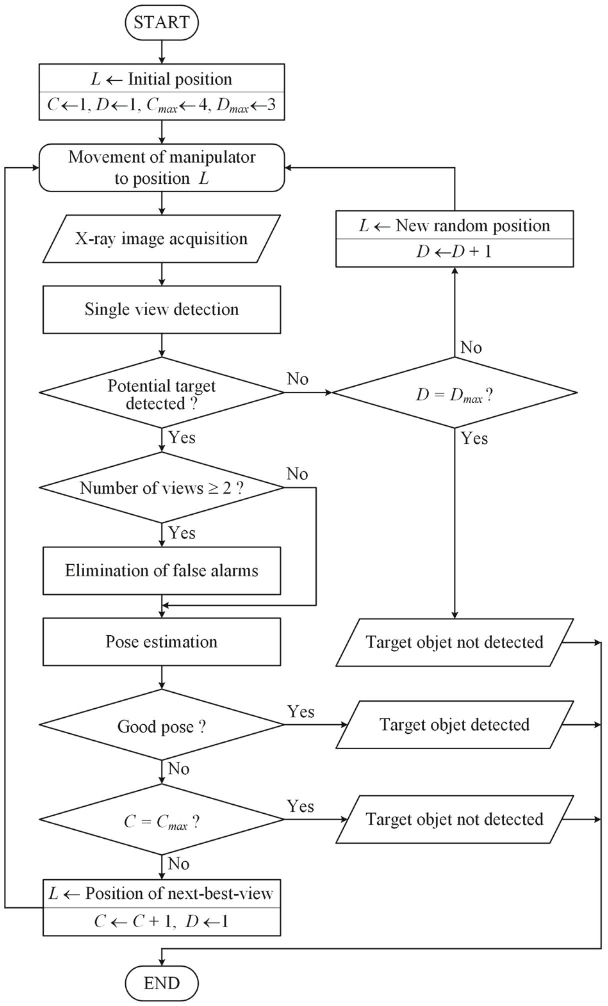


**Fig. 2** Proposed method

includes: detection in single views (Sect. 3.1); pose estimation (Sect. 3.2); next-best-view estimation (Sect. 3.3); and elimination of false alarms (Sect. 3.4). In the following sections, these methods will be explained in further details.

### 3.1 Single-View Detection

In Fig. 2 we can observe the "detector" box, which attempts to detect threat objects from a single view. We use the Adapted Implicit Shape Model (AISM) [2] for object recognition in baggage screening. The two main stages of AISM are:

- *Learning:* The training stage is based on the creation of a visual vocabulary using keypoints and local visual descriptors. In this stage, a target object is represented using a visual vocabulary of parts (category-specific appearance codebook). Keypoints and their local visual descriptors are extracted automatically from all training images of the target object using the recognized SIFT approach [31]. Thus, an object category is characterized by estimating a visual vocabulary of the object parts together with a measurement of their spatial distribution.
- *Testing:* During the testing stage, target objects are detected by searching similar visual words and similar spatial distributions.

More details can be found in [2].

### 3.2 Pose Estimation

For pose estimation, we use the original algorithm that we proposed in [1]. For the sake of completeness, a summary is presented in this section.

During an off-line stage, an X-ray image is taken from every relevant pose of the target object as illustrated in Fig. 3. For each pose $k$, SIFT descriptors $\mathbf{f}_k$ are computed [31]. During the testing stage, a set of SIFT descriptors $\hat{\mathbf{f}}$ of the detected target object is extracted (see Sect. 3.1). Thus, the estimated pose of the detected target object is obtained by:

$$k^* = \underset{k}{\operatorname{argmin}} \left( d(\mathbf{f}_k, \hat{\mathbf{f}}) \right), \tag{1}$$

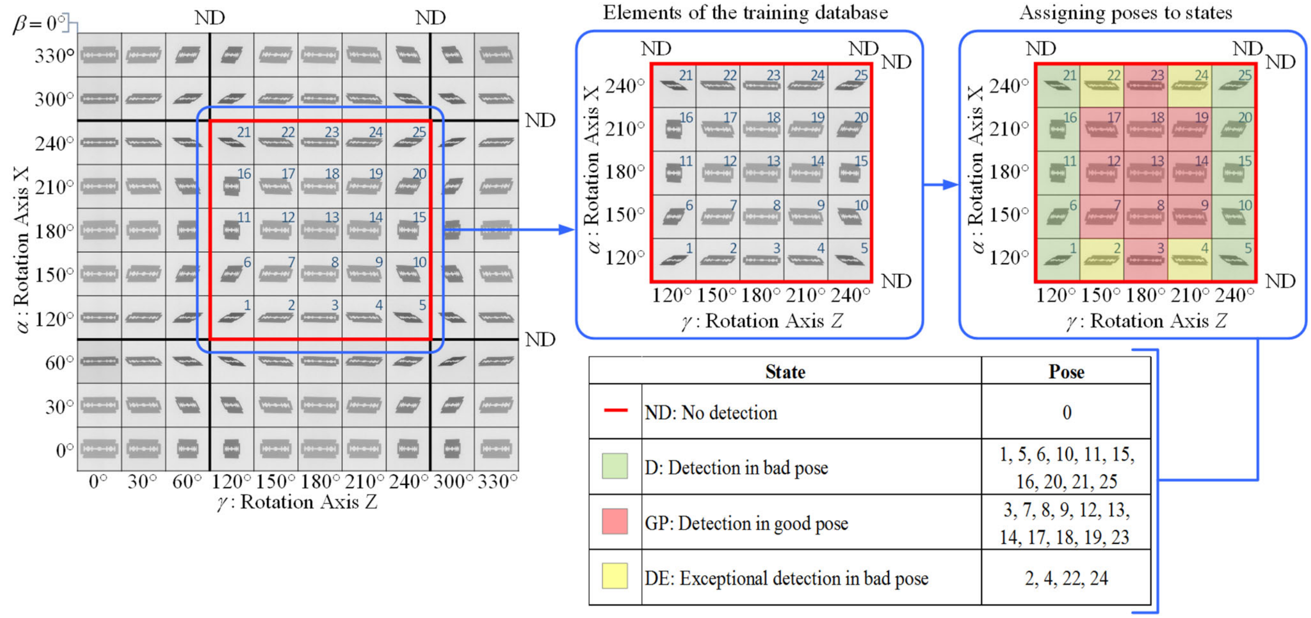
where $d$ is a metrics that measures the difference between the set of SIFT descriptors of pose $k$ and SIFT descriptors of the detected target object. If the distance $d(\mathbf{f}_{k^*}, \hat{\mathbf{f}})$ is not high enough then no pose is estimated. In case of non-detection, the pose estimation algorithm provides a default pose, with a value equal to zero (as shown in the Fig. 3).

In case the target object is detected in a good pose (see Fig. 2), our algorithm decides that no additional view is required for the inspection. However, if a target object is not detected in a good pose (and we have analyzed less than $C_{max}$ X-ray images), then we take another X-ray image using the following algorithm that estimates the next-best-view.

### 3.3 Next-Best-View Estimation

As explained in the previous section, this step is performed only in case an additional view of the test object is required. Thus, the key idea of the active vision strategy is to move the manipulator from its actual position to a new one in which the new pose of the target object allows for its recognition. This new pose should correspond to one of the good poses. These are defined off-line as the poses in which the recognition of the target objects is most likely to occur (Fig. 3). Consequently, an attempt is made to estimate the *next-best-view*. The manipulator can be moved to this position in only one step. In some cases, however, a good pose is achieved after more than one step.

The estimation of the next-best-view is based on the reinforcement learning algorithm 'Q-learning'. This was



| | State | Pose |
|---|---|---|
| — | ND: No detection | 0 |
| 🟩 | D: Detection in bad pose | 1, 5, 6, 10, 11, 15, 16, 20, 21, 25 |
| 🟥 | GP: Detection in good pose | 3, 7, 8, 9, 12, 13, 14, 17, 18, 19, 23 |
| 🟨 | DE: Exceptional detection in bad pose | 2, 4, 22, 24 |

**Fig. 3** Process of assigning poses to states, from a sector of the razor blade training database ($\beta = 0°$)

originally proposed in [32] to solve Markov Decision Processes (MDP) using incomplete information. In Q-learning, an agent learns the optimal policy of its history of interactions with the environment, which is a sequence of experiences represented by a tuple $\langle s, a, r, s' \rangle$. In this definition, the agent was in state $s$, performed action $a$, received reward $r$, and is now in state $s'$. In our approach, the idea is to learn an optimal action-selection to find the next-best-view by estimating $Q(s, a)$ defined as the quantity of a state-action combination. This corresponds to the future value $r + \eta\, V(s')$ that the agent received, where $V(s') = \max_{a'} Q(s', a')$ is the real current reward plus the future value estimated with a discount. Parameter $\eta$ is a number between 0 and 1 ($0 \leq \eta \leq 1$) called the *discount factor* and defined as a tradeoff between exploration and exploitation of the learning process. The agent can upgrade their estimation of $Q(s, a)$ as:

$$Q(s, a) \leftarrow (1 - \alpha)\, Q(s, a) + \alpha \left( r + \eta \max_{a'} Q(s', a') \right), \quad (2)$$

where, $\alpha$ is the *learning rate* ($0 \leq \alpha \leq 1$). This presumes that $\alpha$ is fixed; if $\alpha$ is variable, there will be a different count for each state-action. The Q-Learning Eq. (2) for discreet states, is reduced to Eq. (3), due to the learning rate $\alpha = 1$. Furthermore, in order to make the search for future rewards easier, we set the discount factor ($\eta$) to 0.8. Thus, we privileged exploration over the exploitation of the information already known:

$$Q(s, a) \leftarrow R + \eta \max_{a'} Q(s', a'). \quad (3)$$

In the off-line training stage, we used a database of X-ray razor blade and handgun images in different poses. In Fig. 3 the X-ray images for a razor blade are illustrated. Starting from these images we determined different states for Q-Learning training. Due to the different rotations $\alpha$, $\beta$ and $\gamma$, for $X$, $Y$, $Z$ axes respectively, the quadrants have replicas: because of this, we only used images from the main quadrant during the training process, meaning the images between the angles $\alpha$: [120°, 150°, 180°, 210°, 240°] and $\gamma$: [120°, 150°, 180°, 210°, 240°], for the four angles $\beta$: [0°, 30°, 60°, 90°]. Thus we defined a priori which views should be considered "good poses" (GP) and which ones the algorithm should locate (see Fig. 3 for the different poses and states associated with the razor blade).

Starting from the 25 poses associated with each image, including non-detection (ND), as shown in the Fig. 3, we proceeded to model the environment *noncontinuous*; we grouped together the different images with their respective poses, considering the angle $\gamma$ and we then assigned a name, starting from D (detected, D1 and D2), GP (good pose), DE (exceptional detection, DE1 and DE2) and ND when a good pose is not detected. We were able to simplify this representation using a state diagram (see Fig. 4), where we represented each group of images as a node and each rotation $\gamma$ as a connector. In summary, Q-learning training has three main stages:

- *Definition of poses and states:* Starting from the X-ray images database (see Fig. 3), i.e., the images between the angles $\alpha$: [120°, 150°, 180°, 210°, 240°] and $\gamma$: [120°, 150°, 180°, 210°, 240°], for the four angles $\beta$: [0°, 30°, 60°, 90°], and using a priori knowledge, we defined good

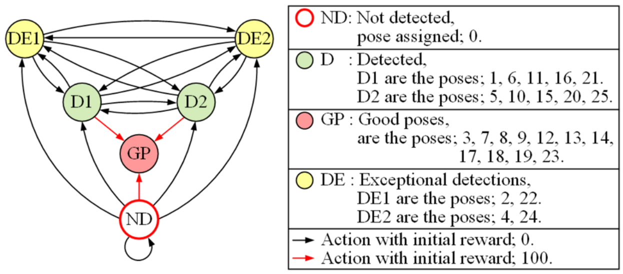

DE1
DE2
D1
D2
GP
ND
○ ND: Not detected, pose assigned; 0.
○ D : Detected, D1 are the poses; 1, 6, 11, 16, 21. D2 are the poses; 5, 10, 15, 20, 25.
○ GP : Good poses, are the poses; 3, 7, 8, 9, 12, 13, 14, 17, 18, 19, 23.
○ DE : Exceptional detections, DE1 are the poses; 2, 22. DE2 are the poses; 4, 24.
→ Action with initial reward; 0.
→ Action with initial reward; 100.

**Fig. 4** States diagram for Q-learning

**Table 1** Learned $Q$ matrix

| | | State *s*' | | | | | | |
|---|---|---|---|---|---|---|---|---|
| State *s* | | DE1 | D1 | DE2 | D2 | ND | GP | Max |
| DE1 | → | 0 | 80 | 64 | 80 | 0 | 0 | D1 |
| D1 | → | 64 | 0 | 64 | 80 | 0 | 100 | GP |
| DE2 | → | 64 | 80 | 0 | 80 | 0 | 0 | D2 |
| D2 | → | 64 | 80 | 64 | 0 | 0 | 100 | GP |
| ND | → | 64 | 80 | 64 | 80 | 80 | 100 | GP |
| GP | → | 0 | 0 | 0 | 0 | 0 | 0 | – |

**Table 2** Set of actions (in $Z$-axis)

| *s* | → | *s*' | $\gamma$ | Description |
|---|---|---|---|---|
| DE1 | → | D1 | −30° | Short negative movement |
| D1 | → | GP | +60° | Long positive movement |
| DE2 | → | D2 | +30° | Short positive movement |
| D2 | → | GP | −60° | Long negative movement |
| ND | → | GP | +60° | Long positive movement |

and bad poses for the target object and generated a states diagram, as shown in Fig. 4.

- *Definition of R Matrix:* $R$ matrix is the data array in which the rewards and punishments ($r$) are stored. The agent receives this matrix from the environment. Rewards indicate to the agent that the procedure is correct, while the punishments indicate that the procedure is incorrect. Rewards are defined as the numeric value that the agent receives after the transition from one state to another ($s \rightarrow s$'). The maximal reward value is +100. This is assigned to an $R$ matrix after a transition to GP. Conversely, a minimal value (zero) as punishment, is assigned to an $R$ matrix if the transition takes place to a pose that is not GP.
- *Definition of Q Matrix:* $Q$ Matrix is the data array in which the corresponding values of the *action-state* pair are stored. In this way, we can search for the optimal route following the highest scores, for the different transitions from a starting state towards a targeted state. The $Q$ matrix is set initially to zero. In each iteration, the $Q$ matrix is updated until convergence.

Once the $Q$ matrix is learned in the training process, we obtain the optimal route to achieve one of the GP. The idea is to follow the links, step-by-step, with the highest values at each state given by the $Q$ matrix. In our case, the learned $Q$ matrix is given in Table 1.

Finally, we estimated the coordinates of the next-best-view using Q-Learning. Thus, we determined an optimal route, i.e., the rotation $\gamma$ around the $Z$ axis that the gripping and rotation system should make to reach a GP of the object. For one transition from state $s$ to state $s$' ($s \rightarrow s$'), our set of actions was made up by one of the four different movements, associated with the state diagram shown in Fig. 4, which were enough to pass through all possible states. These actions are shown in Table 2.

We can illustrate the proposed method with an example. In our example, the threat object was initially detected in pose 2. According to Figs. 3 and 4, this pose corresponds to exceptional detection state DE1. Thus, the next-best-view is estimated as follows: In state $s$ = DE1 (see row DE1 in Table 1), there are two maximum values: D1 and D2 with 80. We arbitrarily choose to move to state D1 (DE1 → D1), our estimated next-best-view. For this action—according to Table 2—, a short negative movement around the Z axis of $\gamma = -30°$ is required. Suppose now, that our algorithm estimates that the new pose is number 6 that corresponds to $s$ = D1. In this case the maximum $Q$ value (see row D1 in Table 1) is given by the GP column with 100. According to Table 2, for D1 → GP, we need to perform a long positive movement in the Z axis at an angle of $\gamma = +60°$. This is our estimated next-best-view. In this example, after two steps the manipulator has reached a good pose using the route DE1 → D1 → GP.

### 3.4 Elimination of False Alarms

Using the explained approach, it is probable that some false alarms occur. In the case that more than one X-ray image of the object being tested is acquired, the elimination of false alarms is possible using multiple view constraints [33], as a threat object that is detected in one image should be viewed from other images as well. Thus, a detection that does not find any correspondence in another view will be considered as a false alarm and will be filtered out. In order to establish the geometric constraints a geometric model is used.

The X-ray image of an object corresponds to a projection in perspective, in which a 3D point of the target object is viewed as a 2D pixel in the digital image of the X-ray, as shown in Fig. 5. The X-ray imaging system consists of an X-ray detector, an X-ray source and a robotic manipulator (rotation system). The geometric model of the X-ray imaging system allows a relationship to be obtained between a 3D

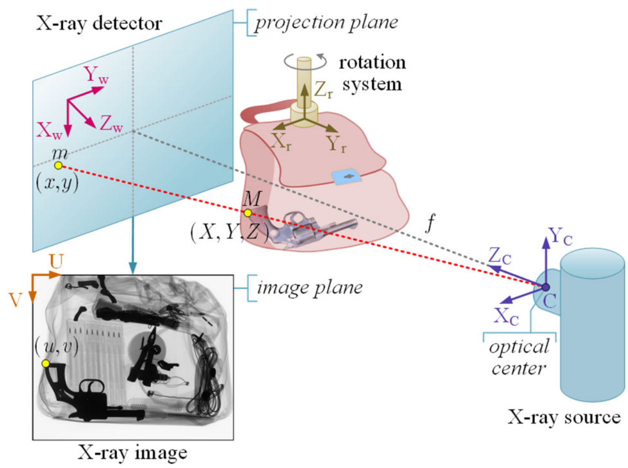


**Fig. 5** Geometric model of an X-ray inspection system

point and its projection in an image as a 2D point. A detailed analysis of the coordinates of an X-ray inspection system can be found in [13]. The geometric model that projects a 3D point $(X, Y, Z)$ into a 2D pixel $(u, v)$ is given in homogeneous coordinates by:

$$\lambda \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{P}_i \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{4}$$

where, $\lambda$ is a scale factor and $\mathbf{P}_i$ is the projection matrix of view $i$. This consists of a $3 \times 4$ element matrix that depends on scale factors, and rotation and translation variables of position $i$. These can be estimated using a calibration approach [34].

Using the geometric model, we are able to establish constraints in multiple views. Thus, detections across the multiple views are validated if they satisfy the geometric criteria as follows:

- *Validation in two views:* If at least two X-ray images are available (images $i$ and $j$), in which the target object has been detected in each one, the detection in both views is validated if the *epipolar* constraint is fulfilled (see Fig. 6) [13]:

$$\frac{|\mathbf{m}_j^\top \mathcal{F}_{ij} \mathbf{m}_i|}{\sqrt{a_1^2 + a_2^2}} < d_0, \tag{5}$$

where, $\mathbf{m}_i$ and $\mathbf{m}_j$ are the centers of mass of the detected objects in each view, $\mathcal{F}_{ij}$ is the fundamental matrix computed from projection matrices $\mathbf{P}_i$ and $\mathbf{P}_j$ from views $i$ and $j$, and $a_1$ and $a_2$ are the coefficients of the $l_j$ epipolar line defined by $l_j = \mathcal{F}_{ij}\mathbf{m}_j = [a_1\ a_2\ a_3]^\mathsf{T}$.

- *Validation in three views:* If at least three X-ray images are available, in which the target object has been detected in each one, detection in the three views can be validated if the trifocal constraint is fulfilled (see Fig. 6) [13]:

$$\|\mathbf{m}_k - \hat{\mathbf{m}}_k\| < d_1, \tag{6}$$

where, $\mathbf{m}_k$ is the center of mass of the detection in the third image and $\hat{\mathbf{m}}_k$ is the estimated position of the hypothetical detection. The last one, $\mathbf{m}_k$, is estimated using the corresponding points $\mathbf{m}_i$ and $\mathbf{m}_j$, and the trifocal tensors $\mathcal{T}_{ij}^k$ computed from projection matrices $\mathbf{P}_i$, $\mathbf{P}_j$ and $\mathbf{P}_k$ of views $i$, $j$ and $k$.

Those detected objects that do not satisfy the multiple view constraints are considered as false alarms, and they will be filtered out.

## 4 Experiments and Results

In this section, we report on the experiments we performed and the results we obtained in the detection of threat objects using the proposed active vision strategy (following Fig. 2). The threat objects we used in our experiments were handguns and razor blades (see Fig. 7). In our approach, we had to train and adjust the test parameters of the single-view detector (AISM detector as explained in Sect. 3.1) for both threat objects. The training consists of the characterization of the target object, which has three steps: (1) training image acquisition: acquisition of representative X-ray images of the threat object; (2) codebook generation: creation of a visual vocabulary using keypoints and local visual descriptors; and (3) occurrence: position estimation of the keypoints related to each visual word of the vocabulary. The adjustment of the test parameters consists of tuning different parameters in the four main stages of object detection: (1) feature extraction; (2) matched codebook entries and voting space; (3) merger of detected candidates; and (4) detection. In our experiments, we used the same parameters reported in our previous work [2] that deals with the single-view detection using AISM.

The performance of our method is measured using the quality evaluation PASCAL criteria,[2] where a detection is considered valid if the normalized area of overlap $a_o$, between the bounding box of a detection $BB_{dt}$ and the bounding box of the ground truth $BB_{gt}$ is greater than a threshold $\theta$. The normalized area is defined as follows (see Fig. 8):

$$a_o = \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})}, \tag{7}$$

[2] From 'PASCAL Visual Object Classes Challenge' [35].

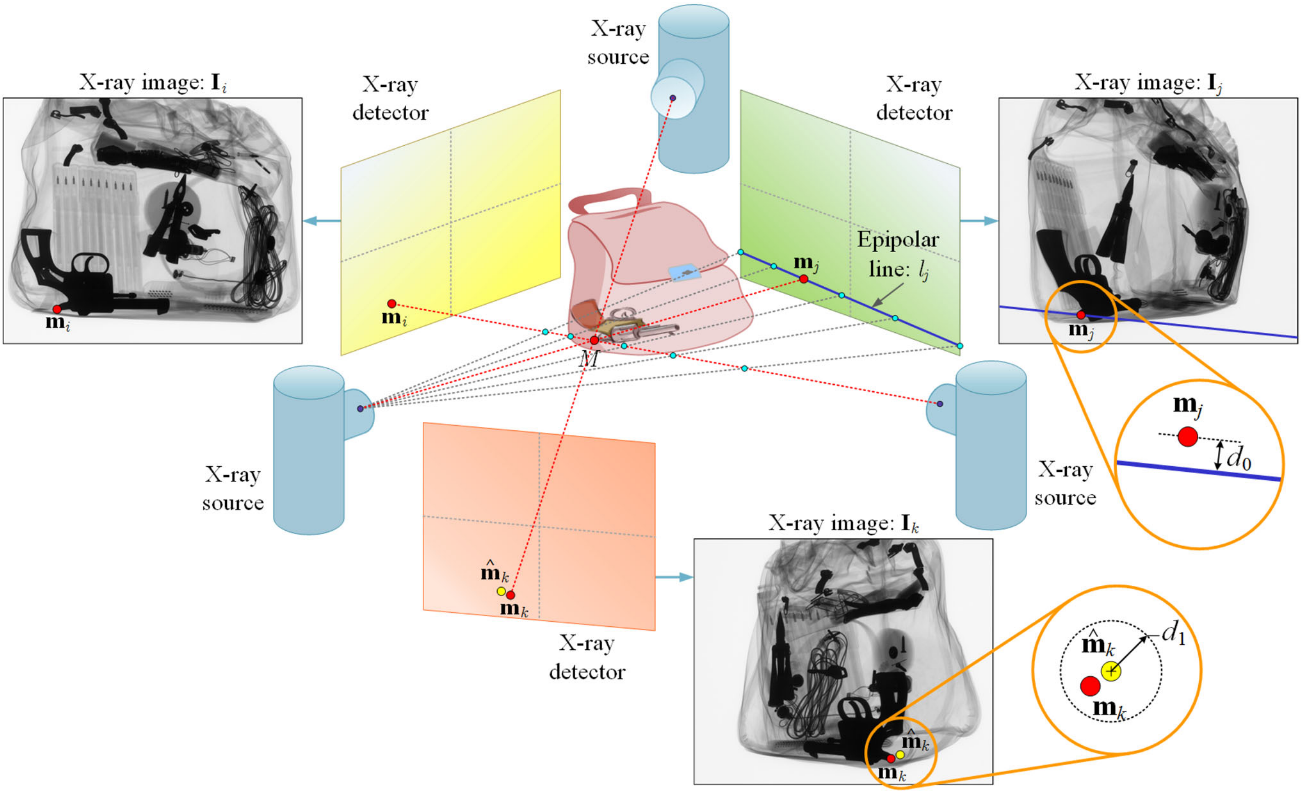


**Fig. 6** Epipolar and trifocal geometry to establish correspondence between points $\mathbf{m}_i$, $\mathbf{m}_j$ and $\mathbf{m}_k$

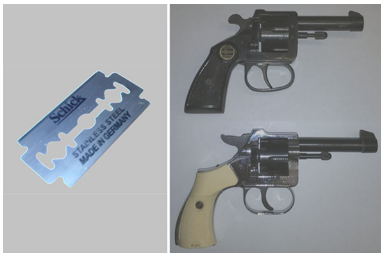


**Fig. 7** Threat objects inside baggage: Razor blade and handguns

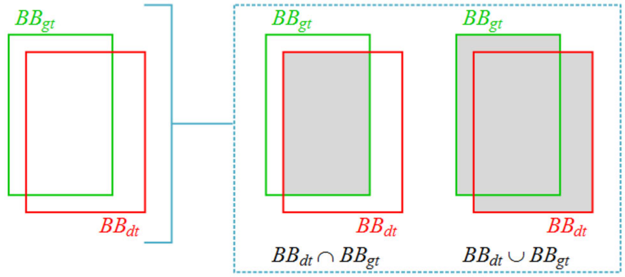


**Fig. 8** Evaluation criteria for comparing bounding boxes. Interpreting the area of overlap criteria. The normalized area $a_o$ is given by the ratio of the intersection to the union areas

where, $BB_{dt} \cap BB_{gt}$ is the intersection of the detection window and the ground truth, and $BB_{dt} \cup BB_{gt}$ their union. In case $a_o > \theta$, the detection is considered true positive, or otherwise false positive. In our work, we measured the precision $(Pr)$ recall $(Re)$ and $F_1$-score defined by:

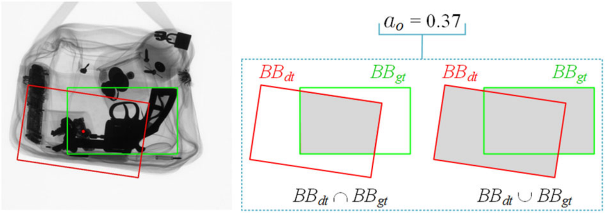


**Fig. 9** Handgun detection with $a_o = 0.37$

$$Re = \frac{\text{TP}}{N_p}, \quad Pr = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad F_1 = 2 \cdot \frac{Pr \cdot Re}{Pr + Re}, \tag{8}$$

where, TP is the number of true positives, FP is the number of false positives and $N_p$ is the total number of objects to be detected. A perfect detector achieves $Pr = 1$ and $Re = 1$, i.e., all objects are detected with no false alarm. In this case, $F_1 = 1$.

In Fig. 9, we show a detection with an overlap area $a_o = 0.37$, and we clearly see that the detection considers almost 2/3 of the visible area of the handgun. Usually, the overlapping threshold $\theta$ is set to 0.5. However, if $\theta$ is 0.5 (or even 0.4), the detection of Fig. 9 would be unfairly con-

**Table 3** Performance for handgun detection

| Method | MV | G | $a_o > 0.3$ | | | $a_o > 0.4$ | | | $a_o > 0.5$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $Pr$ | $Re$ | $F_1$ | $Pr$ | $Re$ | $F_1$ | $Pr$ | $Re$ | $F_1$ |
| Single view | – | – | 0.3300 | 0.1850 | 0.2371 | 0.2650 | 0.1550 | 0.1956 | 0.0750 | 0.0450 | 0.0563 |
| Q-active | Q | – | 0.8250 | 0.6650 | 0.7364 | 0.6050 | 0.4900 | 0.5415 | 0.1750 | 0.1400 | 0.1556 |
| R-active | R | – | 0.5200 | 0.4150 | 0.4616 | 0.4300 | 0.2950 | 0.3499 | 0.1150 | 0.0950 | 0.1040 |

MV: multiple-view strategy, Q: Q-learning, R: random, G: geometric constraints

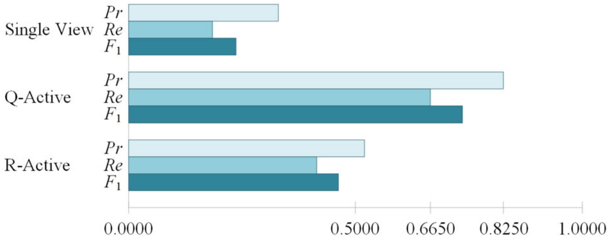


**Fig. 10** Performance for detection of handguns for $a_o > 0.3$ (see details in Table 3)

sidered as false positive. For this reason, in our experiments, we evaluated the performance for $\theta = 0.3$, 0.4 and 0.5.

### 4.1 Detection of Handguns

For the evaluation of our approach in the detection of handguns, we performed 360 experiments as follows: we took two bags, each one containing a handgun and other different objects (see assorted images in Fig. 11a), we rotated each bag 180 times around its $Z$-axis using a robotic manipulator (ABB-Flexpicker) in 2° steps, we acquired an X-ray image in each position yielding two sequences of 180 X-ray images each. For each X-ray image we ran the detection algorithm, and finally we measured the performance of the detection according to (8).

For the baseline method, we tested a single-view detection algorithm (with no multiple-view strategy and no geometric constraint). That means, the AISM-detection algorithm (see Sect. 3.1) made a decision using only one view. As we can see in Table 3 and Fig. 10 (see 'single-view' method) the results are very poor, given that for $\theta = 0.3$, $F_1$-score was only 0.2371 (with $Pr = 0.33004$ and $Re = 0.1850$).

On the other hand, the proposed method using active vision with Q-learning was able to improve the single-view performance significantly achieving $F_1 = 0.7364$ for the same $\theta$ (see 'Q-active' method). In this experiment, we used the same single-view detector as the baseline method; however, we also included the active vision strategy. The single-view detector could recognize only 18.5% of the threat objects ($Re = 0.1850$). That meant that our algorithm had to deal with the poor performance of the detection process in the first view. Active vision could identify better views in order to improve the detection probability. See the example in Fig. 11a where the handgun was detected after three views.

In order to determine the effectiveness of the Q-learning approach, in the estimation of the next-best-view (explained in Sect. 3.3) we replaced the Q-learning estimator with a 'random estimator'. That meant that the next-best-view was a new random position. If the target object is not detected in a GP, the 'random estimator' will output a rotation value randomly selected from the Table 2, this is because such movements are the most accurate to pass from one state to another. The results are summarized in Table 3 and Fig. 10 (see 'R-active' method), where for $\theta = 0.3$, $F_1$-score was only 0.4616.

In these experiments, the algorithm of false alarms elimination (explained in Sect. 3.4) failed. This algorithm is based on geometric constraints that validate the location of corresponding points in different views. In the case of handguns, it was difficult to establish the same representative point of the threat object across the multiple views. Due to the particular asymmetry of the shape of the handgun and the deficiency of the AISM detector for placing the centroid of the bounding box detection $BB_{dt}$ (which is very different from the center of mass of the object), the epipolar and trifocal constraints could not correctly establish the correspondence between the views. For this reason, in the detection of handguns, the algorithm does not consider the step of false alarms elimination. Nevertheless, this step could be used successfully in the detection of razor blades.

### 4.2 Detection of Razor Blades

For the evaluation of our approach in the detection of razor blades, we followed the same methodology explained in the previous section: 360 experiments using two bags that contain the threat object (razor blades). This meant that we had for these experiments two sequences of 180 X-ray images each, that were acquired using a manipulator in steps of 2° degrees around the $Z$-axis of the bags. The reader can see some examples in Fig. 11b.

In this case, the baseline method, i.e., the single-view detection algorithm (with no multiple-view strategy and no geometric constraint), yielded a poor performance as shown in Table 4 and Fig. 11 (see 'single-view' method):

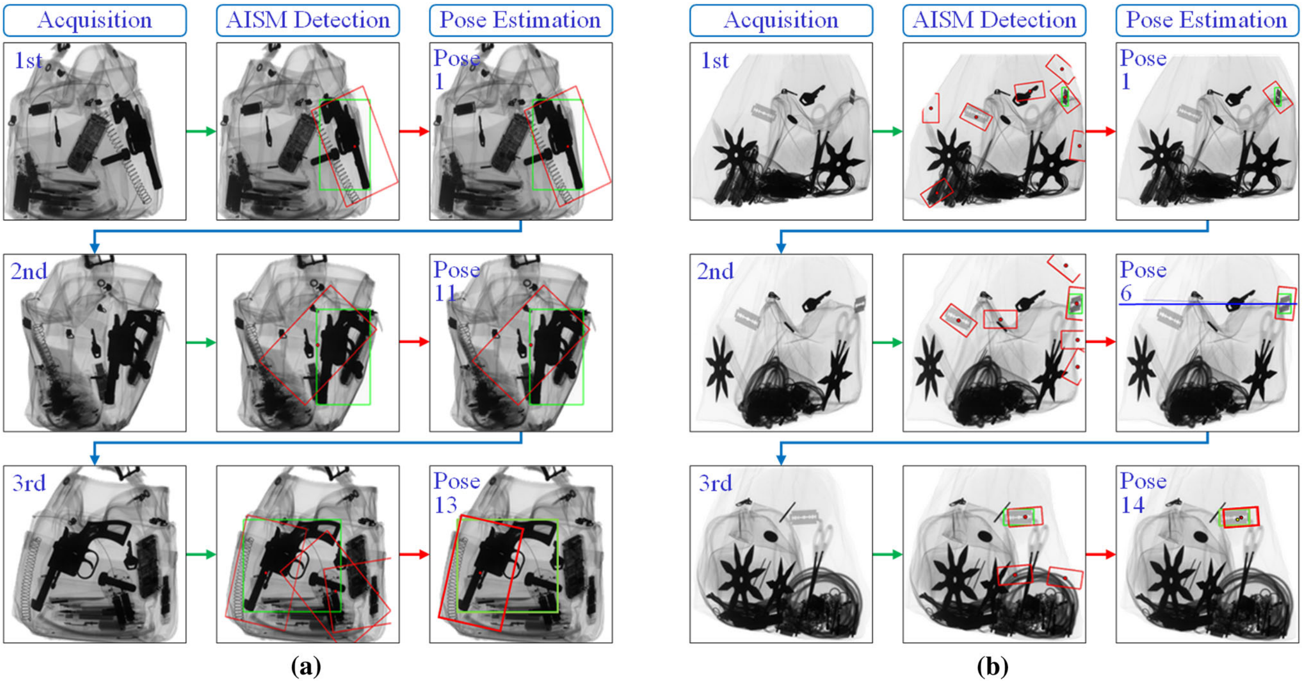


**Fig. 11** Inspection sequences of threat objects, using our active vision approach: **a** handgun detection from pose 1 (bad pose) to pose 13 (good pose), and **b** razor blade detection from pose 1 (bad pose) to pose 14 (good pose)

**Table 4** Performance for razor blade detection

| Method | MV | G | $a_o > 0.3$ | | | $a_o > 0.4$ | | | $a_o > 0.5$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $Pr$ | $Re$ | $F_1$ | $Pr$ | $Re$ | $F_1$ | $Pr$ | $Re$ | $F_1$ |
| Single view | – | – | 0.4750 | 0.3250 | 0.3859 | 0.4650 | 0.3200 | 0.3791 | 0.4350 | 0.3000 | 0.3551 |
| Q-active | Q | – | 0.5550 | 0.5300 | 0.5422 | 0.5550 | 0.5300 | 0.5422 | 0.5200 | 0.4950 | 0.5072 |
| Q-active$_G$ | Q | ✓ | 1.0000 | 0.8150 | 0.8981 | 1.0000 | 0.8150 | 0.8981 | 0.9500 | 0.7700 | 0.8506 |
| R-active$_G$ | R | ✓ | 0.9150 | 0.6400 | 0.7532 | 0.9150 | 0.6400 | 0.7532 | 0.8500 | 0.6100 | 0.7103 |

for $\theta = 0.3$, $F_1$-score was only 0.3859 (with $Pr = 0.4750$ and $Re = 0.3250$).

The proposed method using active vision with Q-learning was able to improve the performance to $F_1 = 0.5422$ (see 'Q-active'), in cases where we did not use the false alarms elimination module. However, when the module of false alarms elimination was used (see 'Q-active$_G$') $F_1$-score increased up to 0.8981. This result highlights the relevance of the use of geometric constraints in active vision. An example of this detection is illustrated in Fig. 11b, where the reader can see the epipolar line in the second image (see the image of Pose 6) and the estimated centroid detection in the third view using trifocal geometry (see the small yellow circle in the image of Pose 14). It is worth mentioning that the module of false alarms elimination did not only reduce the number of false alarms (precision value increased from 0.555 to 1.000, meaning that no false false alarm was detected), but also increased the number of true positives (in which the recall value increased from 0.5300 to 0.8150). The reason for this interesting result is because after two X-ray images, the geometric constraints increased the probability to track the correct threat object. Without geometric constraints, the choice of which potential threat object is to be tracked may fail and for this reason the final detection might be wrong. However, with geometric constraints the tracking is validated in the first two views and the probability to track a real threat object is increased.

In Fig. 11b we show a sequence that highlights the effectiveness of the proposed approach. Here we see how from a bad pose it is possible to reach a GP, starting from a first, second and even a third acquisition of X-ray images. The movements of the robotic manipulator, estimated for the Q-learning algorithm, seem to be mostly adequate, and most complications arise from the internal disorder in the inspected bags, which has a negative influence on the performance of the AISM detector, causing false alarms that are difficult to eliminate, and consequently a diminishing of the global performance of the improved approach.

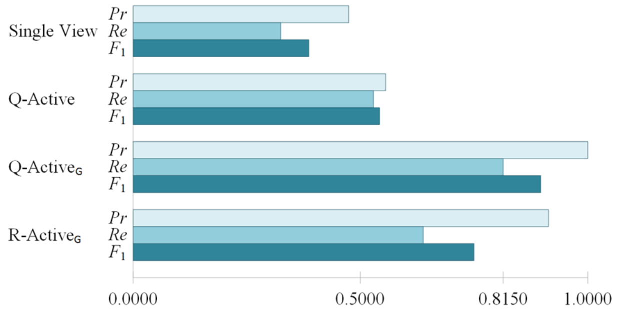


**Fig. 12** Performance for detection of razor blades for $a_o > 0.3$ (see details in Table 4)

Finally, in order to determine the effectiveness of the Q-learning approach with geometric constraints, in the estimation of next-best-view (explained in Sect. 3.3) we replaced the Q-learning estimator by a 'random estimator' as explained in the previous section. The results are summarized in Table 4 and Fig. 12 (see 'R-active' method), where for $\theta = 0.3$, $F_1$-score decreases from 0.8981 to 0.7532.

### 4.3 Implementation Details

In the implementation of our method, we used open source libraries such as VLFeat [36] for $k$-means and SIFT descriptor to implement the AISM detector. The computing time depends on the size and speed of image acquisition, the number of useful descriptors in the image, the spatial distribution, number of occurrences, and rotation speed of the robotic manipulator, among other factors. However, as a reference, the testing results for the detection in a good pose of razor blades were obtained on average after no more than 70 s, and for handgun detection, after no more than 80 s, with respect to each inspection carried out. All the experiments were performed on a Mac Mini Server OS X 10.10.1, processor 2.6 GHz Intel Core i7 with 4 cores and a memory of 16GB RAM 1600 MHz DDR3. The algorithms were implemented in MATLAB 2015a. The code of the MATLAB implementation and all images used in our experiments are available on our webpage.[3]

The X-ray images of our experiments were acquired using a digital X-ray detector (Canon, model CXDI-50G), an X-ray emitter tube (Poskom, model PXM-20BT) and a lead security cabinet to isolate the inspection environment. The size of the X-ray images was 2208 × 2688 pixels. Additionally, for the grip of the bags and rotating movements a robotic manipulator (ABB, model Flexpicker) was used.

## 5 Conclusions

Active vision is not a new concept in computer vision, as it has been used in robotics and throughout industry. Nevertheless, it can be considered a recent strategy in terms of X-ray testing. With this work we have expanded the original approach that we presented in [1], providing a robust method against the internal clutter of the test object and partial occlusion, through the use of: (i) the AISM detector, (ii) the development and incorporation of the estimator of the movement coordinates algorithm to search for a good view, and (iii) the elimination of false alarms using geometric constraints.

This improved approach was developed for the detection of handguns and razor blades placed in bags, and with high degrees of complexity (internal clutter, large quantity of objects and high occlusion levels), meaning more realistic scenarios than those considered in our original approach.

Active inspection allowed us to detect threat objects that are found in intricate or non-representative poses. Our conclusions are threefold:

- *Multiple views represents a good choice in baggage screening:* We can validate the well-known conclusion that strategies based on multiple views achieve higher performance than those based on single views. In our experiments, with handguns and razor blades, the $F_1$-score in single-view experiments was 0.24 and 0.39, respectively. On the other hand, with multiple views, the $F_1$-score was increased by approximately 0.5 in each case. The reason why multiple views are better in baggage screening is because many target objects cannot be recognized by just some (single) views. Such is the case of handguns and razor blades.
- *Q-learning is efficient and effective in X-ray active vision:* the Q-learning algorithm allowed us to estimate the movement coordinates of the next-best-view. This was tested in complex environments of radioscopic inspection, in which there were many levels of uncertainty. The uncertainties that we incorporated into the Q-learning were produced by: (a) the imperfections of the object detector when detecting and estimating good poses, (b) movement restriction of the robotic manipulator, that made it impossible to reach some good poses, and (c) the replicas of the quadrants in the X-ray images database for training, that confused the pose estimator. Thus, the estimator of movement coordinates based on Q-learning is capable of finding an optimal policy of movement that minimizes the quantity of images necessary to locate the target object in a good pose. If we compare a random strategy with a Q-learning strategy, the increment in the performance was significant: $F_1$-score was from 0.46 to 0.74 for handguns and from 0.75 to 0.90 for razor blades.

[3] See http://dmery.ing.puc.cl/index.php/material/.

- *False alarms in baggage screening can be eliminated using geometric constraints:* Geometric multiple view constraints were tested in the detection of razor blades. This is an easy way to improve the performance in multiple views by filtering out those detections that do not find any correspondence in other views. In the active inspection of razor blades, the improvement was from $F_1 = 0.54$ to $F_1 = 0.90$.

Our preliminary experiments have shown that the proposed approach based on active vision achieves promising results. We believe that our method can be employed to aid a user during the task of inspection.

**Acknowledgements** This work was supported in part by DIUDA Grant No. 22277 from Universidad de Atacama, and in part by Fondecyt Grant No. 1161314 from CONICYT, Chile.

## References

1. Riffo, V., Mery, D.: Active X-ray testing of complex objects. Insight Non-Destr. Test. Cond. Monit. **54**(1), 28–35 (2012)
2. Riffo, V., Mery, D.: Automated detection of threat objects using adapted implicit shape model. IEEE Trans. Syst. Man Cybern. **46**(4), 472–482 (2016)
3. Leibe, B., Schiele, B.: Interleaved object categorization and segmentation. In: Proceedings of the British Machine Vision Conference, pp. 78.1–78.10. BMVA Press (2003)
4. Mery, D., Riffo, V., Zscherpel, U., Mondragón, G., Lillo, I., Zuccar, I., Lobel, H., Carrasco, M.: GDXray: the database of X-ray images for nondestructive testing. J. Nondestr. Eval. **34**(4), 42 (2015)
5. Newman, T., Jain, A.: A survey of automated visual inspection. Comput. Vis. Image Underst. **61**(2), 231–262 (1995)
6. Hardmeier, D., Hofer, F., Schwaninger, A.: The role of recurrent cbt for increasing aviation security screeners' visual knowledge and abilities needed in X-ray screening. In: Proceedings of the 4th International Aviation Security Technology Symposium, pp. 338–342, Washington, DC (2006)
7. Schwaninger, A., Hardmeler, D., Hofer, F.: Aviation security screeners visual abilities visual knowledge measurement. IEEE Aerosp. Electron. Syst. Mag. **20**(6), 29–35 (2005)
8. Wales, A., Anderson, C., Jones, K., Schwaninger, A., Horne, J.: Evaluating the two-component inspection model in a simplified luggage search task. Behav. Res. Methods **41**(3), 937 (2009)
9. Bolfing, A., Halbherr, T., Schwaninger, A.: How image based factors and human factors contribute to threat detection performance in X-ray aviation security screening. HCI and Usability for Education and Work. Lecture Notes in Computer Science, vol. 5298, pp. 419–438. Springer, Berlin (2008)
10. Michel, S., Koller, S.M., Schwaninger, A.: Relationship between level of detection performance and amount of recurrent computer-based training. In: Security Technology, 2008. ICCST 2008. 42nd Annual IEEE International Carnahan Conference on, pp. 299–304. IEEE (2008)
11. Schwaninger, A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., Hay, L.: The impact of image based factors and training on threat detection performance in X-ray screening. In: Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008, pp. 317–324 (2008)
12. Zentai, G.: X-ray imaging for homeland security. Imaging systems and techniques. IEEE International Workshop on IST 2008, pp. 1–6 (2008)
13. Mery, D.: Computer Vision for X-Ray Testing. Springer, New York (2015)
14. Mery, D., Svec, E., Arias, M.: Object recognition in X-ray testing using adaptive sparse representations. J. Nondestr. Eval. **35**(3), 45 (2016)
15. Mery, D., Svec, E., Arias, M., Riffo, V., Saavedra, J., Banerjee, S.: Modern computer vision techniques for X-ray testing in baggage inspection. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2016)
16. Uroukov, I., Speller, R.: A preliminary approach to intelligent X-ray imaging for baggage inspection at airports. Signal Process. Res. **4**, 1–11 (2015)
17. Baştan, M., Yousefi, M., Breuel, T.: Visual words on baggage X-ray images. In: Computer Analysis of Images and Patterns. Lecture Notes in Computer Science, vol. 6854, pp. 360–368. Springer, Berlin (2011)
18. Turcsany, D., Mouton, A., Breckon, T.: Improving feature-based object recognition for X-ray baggage security screening using primed visualwords. In: IEEE International Conference on Industrial Technology (ICIT), 2013, pp. 1140–1145 (2013)
19. Zhang, N., Zhu, J.: A study of X-ray machine image local semantic features extraction model based on bag-of-words for airport security. Int. J. Smart Sens. Intell. Syst. **8**(1), 45–64 (2015)
20. Wang, Y., Yang, X., Wu, W., Su, B., Jeon, G.: An X-ray inspection system for illegal object classification based on computer vision. Int. J. Secur. Appl. **10**(10), 155–168 (2016)
21. Akşay, S., Kundegorski, M.E., Devereux, M., Breckon, T.P.: Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 1057–1061 (2016)
22. Mery, D.: Inspection of complex objects using multiple-X-ray views. IEEE/ASME Trans. Mech. **20**(1), 338–347 (2015)
23. Mery, D., Riffo, V., Zuccar, I., Pieringer, C.: Automated X-ray object recognition using an efficient search algorithm in multiple views. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 368–374 (2013)
24. Franzel, T., Schmidt, U., Roth, S.: Object detection in multi-view X-ray images. In: Pattern Recognition. Lecture Notes in Computer Science, vol. 7476, pp. 144–154. Springer, Berlin (2012)
25. Mouton, A., Flitton, G.T., Bizot, S.: An evaluation of image denoising techniques applied to CT baggage screening imagery. In: IEEE International Conference on Industrial Technology (ICIT 2013). IEEE (2013)
26. Flitton, G., Breckon, T.P., Megherbi, N.: A comparison of 3D interest point descriptors with application to airport baggage object detection in complex CT imagery. Pattern Recogn. **46**(9), 2420–2436 (2013)
27. Megherbi, N., Han, J., Breckon, T.P., Flitton, G.T.: A comparison of classification approaches for threat detection in CT based baggage screening. In: 19th IEEE International Conference on Image Processing (ICIP), 2012, pp. 3109–3112. IEEE (2012)
28. Flitton, G., Mouton, A., Breckon, T.P.: Object classification in 3D baggage security computed tomography imagery using visual codebooks. Pattern Recogn. **48**(8), 2489–2499 (2015)
29. Mouton, A., Breckon, T.P.: Materials-based 3D segmentation of unknown objects from dual-energy computed tomography imagery in baggage security screening. Pattern Recogn. **48**(6), 1961–1978 (2015)
30. von Bastian, C., Schwaninger, A., Michel, S.: Do Multi-view X-ray Systems Improve X-ray Image Interpretation in Airport Security Screening?, vol. 52. GRIN Verlag, Munich (2010)

31. Lowe, D.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
32. Watkins, C.J.C.H.: Learning from delayed rewards. Ph.D. thesis, University of Cambridge England (1989)
33. Mery, D., Filbert, D.: Automated flaw detection in aluminum castings based on the tracking of potential defects in a radioscopic image sequence. IEEE Trans. Robot. Autom. **18**(6), 890–901 (2002)
34. Mery, D.: Explicit geometric model of a radioscopic imaging system. NDT E Int. **36**(8), 587–599 (2003)
35. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
36. Vedaldi, A., Fulkerson, B.: VLfeat: an open and portable library of computer vision algorithms. In: MM ’10: Proceedings of the international conference on Multimedia, pp. 1469–1472. New York (2010)