

Assessing X-Ray Image Interpretation Competency of Airport Security Screeners

S. Koller, and A. Schwaninger

Abstract—Baggage screening using x-ray equipment is an essential element of airport security. In order to measure threat detection performance of airport security screeners in a reliable, valid, and standardized way, the X-Ray Competency Assessment Test (X-Ray CAT) has been developed. Based on findings of object recognition studies, X-Ray CAT was designed considering the image-based factors bag complexity, superposition, and viewpoint. Furthermore, it consists of two sets of similar looking threat objects, which allows measuring transfer effects resulting from computer-based training. This study confirmed results obtained in earlier object recognition studies indicating different detection scores for different categories of threat objects and decreasing detection performance with increasing rotation of the threat object. Reliability analyses showed that the X-Ray CAT features high Cronbach Alpha and split half reliability values. Therefore, this test is a useful instrument for initial and recurrent certification and competency assessment of x-ray operators.

Index Terms—Airport security, x-ray screening, screener performance measurement, certification, competency assessment, quality control

I. ASSESSING X-RAY IMAGE INTERPRETATION COMPETENCY USING X-RAY CAT

IN a world of constant risk of terrorist attacks the need of a high level of security in transportation has to be satisfied. One important field is airport security. It is now recognized more and more that well trained aviation security screeners are essential in order to achieve and maintain a high level of security and efficiency at airport security checkpoints. The most expensive equipment is of little value if the humans who operate it are not selected and trained appropriately. Screeners have to know which objects are prohibited and what they look like in x-ray images of passenger bags in order to detect them effectively within a few seconds of inspection time. In order to meet this requirement, several countries use now adaptive computer-based training (CBT) such as X-Ray Tutor (XRT) for initial and recurrent training of screeners ([1], [2]).

A study by [3] showed that recognition of unfamiliar object shapes (e.g., a self-defense gas spray) in x-ray images is poor for non-screeners and much higher for well-trained

aviation security personnel. In addition to such knowledge-based factors also image-based factors such as effect of viewpoint, superposition by other objects, and bag complexity produced by the number and type of other objects ([3], [4]), are considered in XRT. With this CBT screeners can be trained very effectively and efficiently using an individually adaptive algorithm ([5], [6] [7]). XRT contains thousands of x-ray images of bags and of prohibited items depicted in many different viewpoints (see [2] for details). During training, aviation security screeners are exposed to x-ray images of passenger bags containing a prohibited item (threat image) and to harmless bag images. Images are displayed for 10 seconds, and the trainees have to judge for each bag whether it is OK (contains no prohibited item) or NOT OK (contains a prohibited item). XRT is individually adaptive. It starts with threat items depicted in easy views and increases image difficulty by showing threat items in more difficult views in more complex bags and with increasing superposition by other objects. In order to prevent screeners to memorize images of bags, combinations of images of bags and threat objects are created at the point of use. This approach considers the individual training level and visual-cognitive abilities of each screener. It starts with easy images and then increases image difficulty for each individual trainee by showing threat objects in more difficult views in more complex bags and with more superposition by other objects. To conduct periodical measurements of individual screener performance the X-Ray Competency Assessment Test (X-Ray CAT) was developed and integrated in the XRT training system. In X-Ray CAT screeners have to visually search x-ray images of passenger bags for forbidden objects. The visual appearance of x-ray images is the same as during training with XRT. A scientifically reliable and standardized test is essential for individual competency assessment and certification of screeners. Like XRT, the X-Ray CAT has been developed considering scientific findings of threat detection in x-ray images of passenger bags ([3], [4]). As mentioned above, recognition of threat objects in passenger bags is dependent on the viewpoint in which the threat object is seen, the degree of superposition and the degree of clutter in the bag. Superposition is a measure of how much a threat object is superimposed by other objects in the bag (for details see equation below). In X-Ray CAT effects of viewpoint are controlled by using images of threat objects depicted in two standardized rotation angles in easy and difficult view (see below). Images of objects are combined with images of bags in a way that the two views of an object show the same degree of superposition. The bags are chosen such that they

Manuscript received February 26, 2006.

S. Koller and A. Schwaninger are of the Visual Cognition Research Group, Department of General Psychology, University of Zurich, Switzerland (phone: 0041 44 254 38 50; fax: 0041 44 254 38 56; e-mail: a.schwaninger@psychologie.unizh.ch)

are visually comparable concerning the number and form of objects with which they are packed (i.e. the degree of clutter).

The test contains two sets of objects in which object pairs are similar in shape. This allows not only measuring general effects of training by comparing the test results prior and after training, it also provides the possibility to measure transfer effects. By only training the images of one set of the test the threat detection performance for the images of the other set in a next test session indicates whether training of certain objects benefits the recognition of similar other objects. When comparing screener performance on an individual basis, tests must be reliable. Cronbach Alpha and split half reliability analyses were conducted to measure the reliability of X-Ray CAT.

II. METHOD

A. Participants

107 aviation security screeners of a European airport conducted the X-Ray CAT 1.0.0 before starting with the training using XRT SE 2.0.

B. Materials and Procedure

Stimuli were created from Smiths-Heimann Hi-Scan 6040i colour x-ray images of prohibited items and passenger bags (Fig. 1 displays an example of the stimuli). Four categories of prohibited items were chosen based on the categorization of current threat image projection systems: guns, improvised explosive devices (IEDs), knives and other prohibited items (e.g., gas, chemicals, grenades etc.). Of each category 16 exemplars are used (8 pairs). Each pair consists of two prohibited items that are similar in shape. These were divided into two sets, set A and set B. Furthermore, every item is depicted in two different viewpoints. The easy viewpoint shows the object in canonical view (see gun in Fig.1); the difficult viewpoint shows it with an 85 degree horizontal rotation or an 85 degree vertical rotation. In each threat category half of the prohibited items of the difficult viewpoint are rotated vertically, the other half horizontally. Set A and B are equalized in regard to the rotations of prohibited objects.

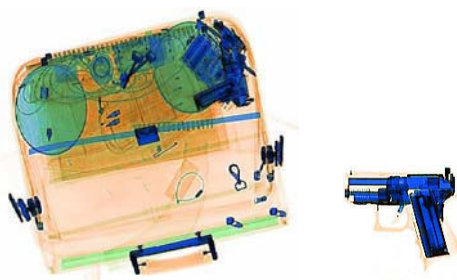


Fig. 1. Example of an x-ray image of a passenger bag containing the prohibited item depicted separately on the bottom right.

Every prohibited item was combined with a bag in a manner that the degree of superposition for one item is the same for both viewpoints. This was achieved using an image processing tool that calculates the difference of the brightness between the pixels of the two superimposed images using the following formula for superposition:

$$SP = \frac{\sqrt{I_{SN}(x,y) - I_N(x,y)}}{Threshold}$$

SP = Superposition; I_{SN} = Grayscale intensity of the SN (Signal plus Noise) image (contains a prohibited item); I_N = Grayscale intensity of the N (Noise) image (contains no prohibited item); Threshold: Number of pixels of the prohibited item where R, G and B are < 253

Using this equation the superposition value is independent of the size of the prohibited item. This value can be kept constant for the two views of a threat object, independent of the degree of clutter in the bags, when combining the bag image and the prohibited item. The bag images were checked by at least two aviation security experts to be sure they do not contain any other prohibited item. Clean bag images were assigned to the different categories and viewpoints of the prohibited items in a way that their image difficulty was balanced across all categories. This was done using the false alarm rate for each image based on a pilot study. Each bag is used twice, once containing a prohibited item (threat image) and once without (harmless image). Thus, the X-Ray CAT is composed of 256 test trials: 4 threat types (guns, IEDs, knives, other) * 8 (exemplars) * 2 (sets) * 2 (views) * 2 (harmless images vs. threat images).

The X-Ray CAT is integrated in the XRT training system and takes about 2-3 sessions of 20 minutes to complete. The visible appearance of the test is the same as during training. The task is to visually inspect the images and to judge whether they are OK (contain no prohibited item) or NOT OK (contain prohibited item). In this study, images disappeared after 10 seconds. In addition to the OK / NOT OK response, screeners had to indicate the perceived difficulty of each image on a 100 point scale (difficulty rating; 1=easy, 100=difficult). All responses are given by pressing buttons on the screen.

III. RESULTS

The two sorts of x-ray images used in the test (bags containing a prohibited item and bags containing no prohibited item) result in four different possible outcomes per trial: hits (correctly found threat objects), misses (missed threat objects), false alarms (incorrectly reporting a threat object) and correct rejections (correctly judged harmless bag as being OK). The hit rate alone is not a valid measure of detection performance in terms of sensitivity [10]. The reason is simple; a test candidate can achieve a high hit rate by simply judging most bags as NOT OK. Therefore, the false alarm rate has to be taken into account, in addition to the hit rate. A “non-parametric” measure of detection performance is A’ which is calculated using the hit and false alarm rates adopting the following formula [8]:

$$A' = 0.5 + [(H - F)(1 + H - F)]/[4H(1 - F)],$$

whereas H is the hit rate and F the false alarm rate. If the false alarm rate is greater than the hit rate the equation must be modified [9]:

$$A' = 0.5 - [(F - H)(1 + F - H)]/[4F(1 - H)].$$

For details and other measures of x-ray screening performance see [10].

The results section provides detection performance, reaction time values, reliability measures and the results of an analysis of variance (ANOVA). The ANOVA was carried out to investigate whether detection performance varies across different threat categories (guns, IEDs, knives, others). In addition, potential differences and interactions regarding the two sets of images were examined.

A. Detection Performance

Fig. 2 shows the detection performance for each of the four categories of threat objects separately and for both views, averaged across all screeners. Detection performance is significantly higher for threat objects shown in canonical view than for threat objects that are rotated 85 degrees. No numeric values are shown since these are security sensitive data.

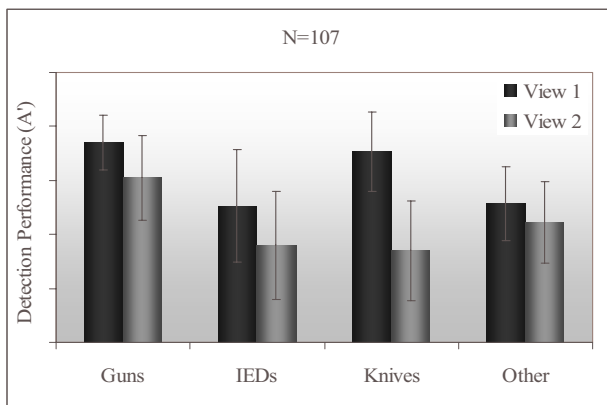


Fig. 2. Detection Performance A' broken up by category and unrotated (View 1) vs. 85° rotated objects (View 2). The thin bars are standard deviations. Pairwise comparisons showed significant viewpoint effects for all four threat categories (all $p < .001$).

B. Reaction Times

For each image, reaction times (RTs) were measured, i.e. the duration from image onset until an answer (OK or NOT OK) was given. The mean reaction times and standard deviations for each category and viewpoint are displayed in Fig. 3. In contrast to the results of detection performance, there were no viewpoint effects for the reaction times.

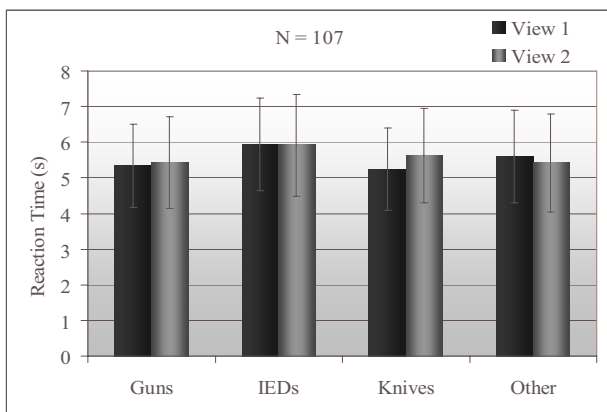


Fig. 3. Reaction times broken up by the four categories and unrotated (View 1) vs. 85° rotated objects (View 2). Thin bars represent standard deviations.

Bivariate Pearson correlations between A', RT and difficulty ratings were calculated. A' correlates with reaction times $r = -.353$ ($p < .001$) and with difficulty ratings $r = -.213$ ($p < .05$).

C. Reliability of the X-Ray CAT

Cronbach Alpha and Guttman split-half reliabilities were calculated based on hits and correct rejections (PC = percentage correct) as well as on difficulty ratings (DR). The analyses were made separately for test trials including a threat item (SN trials) and those not including a threat item (N trials). Table 1 shows the reliability coefficients.

TABLE 1. RELIABILITIES

		RELIABILITY ANALYSES			
Reliability Coefficients		PC SN	PC N	DR SN	DR N
	X-Ray CAT	Alpha	.878	.914	.975
	Split-half	.840	.903	.964	.980

Note. PC = Percent Correct, DR = Difficulty Ratings, SN = Bags containing a threat ("Signal plus Noise Trials"), N = Bags containing no threat ("Noise-Trials")

Additionally, a three-way ANOVA was conducted on A' scores with the three within-participants factors threat category, viewpoint and set. It showed significant main effects of prohibited items category (guns, IEDs, knives and other) with an effect size of $\eta^2 = .52$, $F(3, 318) = 112.61$, $MSE = .014$, $p < .001$, and viewpoint (canonical vs. rotated) $\eta^2 = .76$, $F(1, 106) = 343.39$, $MSE = .010$, $p < .001$, but no main effect of the set. The following two-way interactions were significant: Prohibited items category * viewpoint $\eta^2 = .35$, $F(3, 318) = 55.84$, $MSE = .009$, $p < .001$, prohibited items category * set $\eta^2 = .09$, $F(3, 318) = 10.46$, $MSE = .008$, $p < .001$, and view * set $\eta^2 = .05$, $F(1, 106) = 5.66$, $MSE = .006$, $p < .05$. These results indicate different detection performance for different prohibited items categories and higher detection performance for prohibited items in canonical view than for rotated prohibited items (effect of viewpoint, see also [3], [4]). This is consistent with results reported in the view-based object recognition literature (see [12] and [13] for reviews). The effect sizes were very large according to the conventions by [14]. The factor set shows no main effect and the effect sizes of the two interactions are small to medium ([14]). This indicates that the two sets are comparable in terms of image difficulty. Another fact supporting the comparability of the two sets is the overall correlation of $r = .715$ ($p < .001$) between them.

IV. DISCUSSION

The intention of this study was the development of a reliable test for the measurement of threat detection performance of airport security screeners. The X-Ray CAT contains images of the four prohibited items categories guns, IEDs, knives, and other prohibited items that are depicted in two different viewpoints (not rotated vs. rotated). High Cronbach Alpha coefficients showed that the X-Ray CAT is

a reliable and useful tool for this important purpose (all $\alpha > .88$). Further evidence for the good reliability of this test was revealed by the almost equally high Guttman split-half reliability (all $r > .84$). The two set composition of the test allows the investigation of transfer effects in training studies. An ANOVA on A' indicated no overall statistical difference between the two image sets. The aim of a transfer effect study would be to find out, if after training certain threat objects, a screener can transfer the acquired visual knowledge to similarly looking objects. At the moment this is being investigated in three European countries and in Canada using the X-Ray CAT. The importance of training not only canonical views of prohibited items but also rotated views is emphasized by the main effect of viewpoint in the ANOVA ($\eta^2 = .76, p < .001$). The results support results from earlier studies showing that detection performance can vary substantially depending on object rotation ([3], [4]). Another important result of this study was the fact that detection performance varies substantially depending on the prohibited items category. Detection performance was best for guns, followed by knives, other threat types, and IEDs. This pattern is consistent with the view that that knowledge and experience about which prohibited items exist and what they look like in x-ray images is essential for an aviation security screener to quickly and reliably perform the x-ray screening task.

Although RTs and detection performance A' correlate significantly, there were no viewpoint effects for RTs. This indicates that there is no speed-accuracy trade-off regarding effects of viewpoint. The lower detection performance for rotated objects can not be due to faster and thus less accurate responses of the screeners. The correlation between A' and reaction time values across all threat categories and both viewpoints could represent higher reaction times for more difficult threat objects on average. In other words, screeners need on average more time to respond to an image containing a more difficult threat object. This would correspond to findings in threat object recognition, that more time is spent for searching an image when a prohibited item is not recognized [2]. Search is discontinued as soon as a potential threat object is found and the answer NOT OK can be given. If no such object can be found, search can be prolonged until the image disappears from the monitor. The lower detection performance could result from a higher miss-rate (and thus lower hit rate) which could imply longer search time. This assumption finds confirmation in the significant correlation between the hit rate and average reaction time for each image containing a threat object ($r = -.683, p < .001$). This negative correlation signifies shorter reaction times for higher hit rates that is, the higher the probability of an image to be correctly judged as NOT OK (containing a prohibited item), the less time is needed for finding the threat object.

The significant correlation between A' and difficulty ratings (across all categories and both viewpoints) reveals the tendency of better recognized threat images to be estimated by the screeners as being easier to judge (although this correlation was rather small).

The X-Ray CAT can be used to measure effects of training by comparing the results of the test prior and after

training. This is currently being done in other studies conducted in three European countries and Canada. Since the X-Ray CAT is implemented in the XRT training system, the test taking procedure is very easy. The X-Ray CAT is also a tool for quality control as it allows airports to obtain measures of the detection performance of their screeners, screening companies and checkpoints. These can be compared with each other as well as other airports or over time. Moreover, the high reliability of this test provides a solid basis for using this test in order to measure individual screener x-ray image interpretation competency and to conduct periodical certification.

ACKNOWLEDGMENT

This research was financially supported by Zurich Airport Unique, Switzerland. We are thankful to Zurich State Police, Airport Division for their help in creating the stimuli and the good collaboration for conducting parts of this study.

REFERENCES

- [1] A. Schwaninger, "Computer based training: a powerful tool to the enhancement of human factors," *Aviation Security International*, FEB, pp. 31-36, 2004.
- [2] A. Schwaninger, "Increasing Efficiency in Airport Security Screening," *WIT Transactions on the Built Environment*, vol. 82, pp. 405-416, 2005.
- [3] A. Schwaninger, D. Hardmeier and F. Hofer, "Measuring visual abilities and visual knowledge of aviation security screeners," *IEEE ICCST Proceedings*, vol. 38, pp. 258-264, 2004.
- [4] A. Schwaninger, "Evaluation and selection of airport security screeners," *AIRPORT*, vol. 2, pp. 14-15, 2003.
- [5] A. Schwaninger, "Training of airport security screeners," *AIRPORT*, vol. 5, pp. 11-13, 2003.
- [6] A. Schwaninger, "Computer based training: a powerful tool to the enhancement of human factors," *Aviation Security International*, pp. 31-36, February 2004.
- [7] A. Schwaninger and F. Hofer, "Evaluation of CBT for increasing threat detection performance in X-ray screening," in *The Internet Society 2004, Advances in Learning, Commerce and Security*, K. Morgan and M. J. Spector, Eds. Wessex : WIT Press, 2004, pp. 147-156.
- [8] J.B. Grier, "Nonparametric indexes for sensitivity and bias: Computing formulas," *Psychological Bulletin*, vol. 75, 424-429, 1971
- [9] D. Aaronson and B. Watt, "Extensions of Grier's computational formulas for A' and B'" to below-chance performance," *Psychological Bulletin*, vol. 102, pp. 439-442, 1987
- [10] F. Hofer and A. Schwaninger, "Reliable and valid measures of threat detection performance in X-ray screening," *IEEE ICCST Proceedings*, pp. 303-308, 2004.
- [11] A. Schwaninger, "Reliable measurements of threat detection," *AIRPORT*, vol. 1, pp. 22-23, 2003.
- [12] M. J. Tarr and H. H. Bülthoff, "Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993)," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, pp. 1494-1505, 1995.
- [13] M. J. Tarr and H. H. Bülthoff, "Image-based object recognition in man, monkey and machine," in *Object recognition in man, monkey and machine*, M. J. Tarr and H. H. Bülthoff, Eds. Cambridge, MA: MIT Press, 1998, pp. 1-20.
- [14] J. Cohen, *Statistical power analysis for the behavioral sciences*. New York: Erlbaum, Hillsdale, 1988.