Airport Security X-Ray Screening of Hold Baggage: 2D Versus 3D Imaging and Evaluation of an on-Screen Alarm Resolution Protocol

Conference Paper · October 2018
DOI: 10.1109/CCST.2018.8585713

3 authors:

Nicole Hättenschwiler
University of Applied Sciences and Arts Northwestern Switzerland
14 PUBLICATIONS · 17 CITATIONS

Sarah Merks
University of Applied Sciences and Arts Northwestern Switzerland
6 PUBLICATIONS · 12 CITATIONS

Adrian Schwaninger
University of Applied Sciences and Arts Northwestern Switzerland
145 PUBLICATIONS · 2,043 CITATIONS

Some of the authors of this publication are also working on these related projects:

3D CT X-Ray screening View project

All content following this page was uploaded by Nicole Hättenschwiler on 23 October 2018.
The user has requested enhancement of the downloaded file.
Airport security X-ray screening of hold baggage: 
2D versus 3D imaging and evaluation of an on-screen alarm resolution protocol

Nicole Hättenschwiler, Sarah Merks, and Adrian Schwaninger
School of Applied Psychology
University of Applied Sciences and Arts Northwestern Switzerland (FHNW)
Olten, Switzerland
and
Center for Adaptive Security Research and Applications (CASRA)
Zurich, Switzerland
Email: nicole.haettenschwiler@fhnw.ch

Abstract— In airport security screening, passenger baggage that is transported in the hold of an aircraft (hold baggage) is screened using X-ray machines with explosive detection technology. Older systems are based on 2D multi-view imaging whereas newer systems are based on computer tomography (CT) that features 3D rotatable images (3D imaging). Regulators, airport operators and security providers currently discuss whether extensive and specific training is necessary for screeners who are used to 2D multi-view imaging before they start working with 3D imaging. Moreover, to facilitate the decision making of screeners, so called on-screen alarm resolution protocols (OSARP) are available for 3D imaging. However, their effectiveness has not been investigated yet. To address these issues, we compared the visual inspection performance of screeners using state-of-the-art 2D multi-view imaging versus 3D imaging versus 3D imaging following a specific on-screen alarm resolution protocol (OSARP).

In this study, screeners had to decide whether X-ray images contained an improvised explosive device (IED) or not. Results showed that there was no difference in detection performance ($d'$) between 2D and 3D imaging. Visual inspection with 3D imaging following an OSARP resulted in higher detection performance ($d'$) compared to 2D and 3D imaging. In conclusion, screeners currently working with 2D multi-view technology do not need extensive and specific training to achieve comparable detection performance ($d'$) with 3D imaging.

Keywords— Airport security screening, 3D imaging, X-ray imaging technology, on-screen alarm resolution, visual search

I. INTRODUCTION

Compared to cabin baggage screening, where multiple target types (guns, knives, IEDs, explosives, other threats) pose a threat, there is only one threat category in hold baggage screening. As passengers cannot access items stored in the hold of an aircraft, guns and knives in hold baggage do not pose a threat. Therefore, hold baggage screening targets only functioning improvised explosive devices (IEDs) [1]. At airports, all hold baggage is screened by X-ray machines that are explosive detection systems (EDS). They indicate areas in X-ray images that might be explosive by colored frames or a specific surface color [2]. X-ray images on which an EDS has raised an alarm are sent to remote screening locations for on-screen alarm resolution (OSAR) by airport security officers (screeners). The task of hold baggage screeners is to visually inspect alarmed areas in X-ray images and decide whether such EDS alarms are harmless (false alarms of the EDS) or whether the hold baggage needs further inspection because it might contain an IED. Therefore, in HBS, the screeners’ task is mainly a decision task whereas in cabin baggage screening, visual inspection consists of search and decision making [3], [4].

In order to decide whether a bag contains a fully functioning IED or not, screeners need to identify the following necessary components of an IED: a triggering device, a power source, explosive mass, and a detonator that need to be connected to each other by, for example, wires [2], [5]. Through computer-based training, screeners can learn to recognize these components, and they can achieve and maintain a high detection performance for IEDs [6] – [11]. International regulations consider this by mandating extensive and recurring training of screeners. For example, European regulations mandate at least 6 hours of image recognition training and testing in every 6-month period [12].

To screen hold baggage at airports, 2D multi-view X-ray imaging systems are currently the most widely used technology. However, 2D multi-view systems are not able to unambiguously reveal the exact bag content for complex packed and cluttered bags [13]. Newer technology is based on computer tomography (CT). Such systems generate a stack of contiguous slice images that are used to calculate and visualize a volumetric view of the bag [14]. Screeners have then the possibility to inspect bags as 3D-rotatable images that are enhanced by using depth cues with the additional option to slice through them [2], [14], [15]. Due to these additional functions, 3D-rotatable images should facilitate the recognition of threats when they appear from an unusual viewpoint or if they are superimposed by other objects [16], [17]. Further, continuous exposure to 3D objects could
result in richer visual object representations [18], [19], that could improve screeners' visual inspection performance in general. However, CT systems have lower image quality compared to 2D imaging [13], [15], [20], which could impair detection performance with 3D imaging. It is therefore interesting to compare 2D and 3D imaging because it is unclear whether the benefits of 3D imaging outweigh the potential negative effects of lower image quality.

Comparing 2D and 3D imaging is also of high practical relevance: Regulators, airport operators and security providers currently discuss whether extensive and specific training is necessary for screeners who are used to 2D multi-view imaging before they can be allowed to work with 3D imaging. Therefore, the first goal of this study was to compare the visual inspection performance of 2D screeners using 2D versus 3D imaging. Moreover, to facilitate decision making of screeners, so called on-screen alarm resolution protocols (OSARP) are available for 3D imaging that are assumed to improve screeners inspection performance. As their effectiveness has not been investigated yet, the second goal of this study was to compare 2D screeners' visual inspection performance using 3D imaging with and without following a specific on-screen alarm resolution protocol (OSARP).

To address these issues, we tested 2D screeners using a simulated hold baggage screening task under three conditions: 2D multi-view imaging versus 3D rotatable imaging versus 3D rotatable imaging following OSARP.

II. METHODS

A. Participants

Participants were \( n = 62 \) professional hold baggage screeners from an international airport with work experience using 2D multi-view imaging (for descriptive statistics see Table 1). All screeners had been selected, qualified, trained, and certified according to the standards set by the appropriate national authority (civil aviation administration) in compliance with the relevant EU regulation [12]. The current research complied with the American Psychological Association Code of Ethics and was approved by the Review Board of the School of Applied Psychology of the University of Applied Sciences and Arts Northwestern Switzerland. Informed written consent was obtained from all participants.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description of Screeners Participating in the Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n^a )</td>
</tr>
<tr>
<td>2D</td>
<td>20</td>
</tr>
<tr>
<td>3D</td>
<td>22</td>
</tr>
<tr>
<td>OSARP</td>
<td>20</td>
</tr>
</tbody>
</table>

\( ^a \) Two screeners dropped out due to illness after the first test date. Another screener had to be excluded prior to data analyses due to a system error of a simulator.

B. Design

Each screener first performed a pretest to familiarize with the 2D and 3D simulators and the testing procedure\(^1\). Screeners were assigned to balanced groups to be tested with either 2D imaging, 3D imaging (Fig. 1) or 3D imaging following an OSARP training. All three groups conducted the main test two weeks after the pretest using a between-subjects design with condition (2D vs. 3D vs. OSARP) as independent variable and visual inspection measures as dependent variables.

C. Apparatus

Screeners were tested using simulators provided by the manufacturer of the imaging systems. There were six individual testing stations with 19” TFT monitors and the room was dimly lit for testing. Four to six participants were tested at a time performing the test individually, quietly, and under supervision. This is a typical scenario in hold baggage screening [22].

D. Stimuli

X-ray image recording of baggage was conducted at a test center of a national transportation security organization. IEDs were prepared by an IED expert serving as consultant for this study. Thirty-two different bags were repeatedly used by repacking them to create unique stimuli for the pretest and the main test. All bags were packed in a way that resulted in medium X-ray image complexity as judged by the IED expert and the authors. Target-present images contained one IED and target-absent images one EDS false alarm (e.g., cheese, certain liquids, etc.).

The pretest consisted of 64 bag X-ray images recorded with 2D imaging and 64 different bag X-ray images recorded with 3D imaging. Target prevalence was 50% using 32 IEDs that
were shown twice in different bags using medium superposition: once recorded from a more frontal perspective, and once from a horizontally or vertically rotated perspective. The main test consisted of 256 bags that were recorded with both the 2D and 3D imaging system. To ensure the same system reliability (e.g., [23]), we used EDS alarms from the 3D imaging system as a reference when setting red frames manually around the objects of interest in the 2D imaging stimuli. Target prevalence was 50% using 32 different IEDs than in the pretest. Each IED was used four times in four different bags by varying viewpoint and superposition.

E. Procedure

For the pretest and the main test, screeners were instructed to visually inspect each X-ray image as if they were working at the airport and decide as accurately and quickly as possible whether or not the image contained a target (IED) by clicking on a target-present or a target-absent button on the simulator interface (a yes–no task in signal detection theory; see [24]). After the instructions, all participants started with 10 practice trials (5 target-absent and 5 target-present images in random order). A time limit of 60 s was set for viewing an X-ray image.

Screeners in the OSARP condition received a short training before conducting the main test. The original protocol developed by the IED expert is usually taught in two days and is customized to the images and interface of the 3D CT machine used in this study. This OSARP was adapted and shortened for our study so that it could be taught in 40 min. It included two heuristic steps on how to decide whether the bag is harmless or whether it contains an IED. After the training, the screeners took a break of 15 min before conducting the main test.

As the European regulation mandates that screeners have to take a break of at least 10 min after continuous visual inspection of X-ray images [12], the main test was divided into two blocks and screeners took a break of 10 to 15 min after completing the first block. Block order was counterbalanced across participants. Images appeared in random order within a block. All participants completed the pretest in less than 40 min and the main test in less than 1.5 hr including breaks.

F. Statistical Analyses

We computed analyses of variance (ANOVA) with detection performance (d’), response bias (c), target-absent RT, and target-present RT as dependent variables and test condition (2D, 3D and 3D with OSARP) as independent variable. All ANOVAs were conducted with SPSS version 22 and alpha was set at 0.05. Post hoc comparisons were conducted with R version 3.22 [25] and Holm–Bonferroni corrections were applied [26]. Effect sizes of ANOVAs are reported with ηp² (partial eta-squared); effect sizes of t tests, with Cohen’s d. Our dependent variables detection performance (1) and response bias (2) were calculated using the following signal detection theory (SDT) formulae, whereby z refers to the inverse of the cumulative distribution function of the standard normal distribution [24], [27]:

\[ d' = z(HR) - z(FAR) \]  
\[ c = -0.5 * z(FAR) - z(HR) \]

III. RESULTS

Fig. 2 shows detection performance d’ depending on the three test conditions.

A one-way between subjects ANOVA with detection performance (d’) as dependent variable 2 and condition as independent variable showed a significant main effect, \( F(2, 62) = 3.28, p = 0.045, \eta_p^2 = 0.100 \). To investigate whether detection performance (d’) was different between the 2D and 3D imaging condition, a two-tailed t-test was performed. No statistical difference was found, \( t(40) = 0.145, p = 0.886, d = 0.05 \). To test the hypothesis that OSARP results in better detection performance (d’) than 3D imaging, a one-tailed t-test was conducted. This showed that OSARP increased detection performance of screeners, \( t(40) = 3.05, p = 0.002, d = 1.05 \). Fig. 3 shows response bias c by test conditions.

---

\[ ^2 \text{Sensitivity was recalculated under the assumption of an unequal variance distribution of threat- and false-alarm images [24]. A one-way between subjects ANOVA with da as sensitivity measure (with a slope parameter of 0.6) revealed a significant effect for condition } F(2, 62) = 19.01, p < 0.001 \text{ and significant post-hoc comparisons between 2D and OSARP (} p < 0.001 \text{) and 3D and OSARP (} p < 0.001 \text{).} \]
A one-way between subjects ANOVA with response bias (c) as the dependent variable showed a significant effect for condition, F(2, 62) = 15.69, p < 0.001, \( \eta^2 = 0.35 \). Posthoc tests using Holm-Bonferroni corrections showed that response bias of the OSARP condition was significantly lower and therefore more neutral than of the 2D (p < 0.001) and 3D condition (p < 0.001). The difference between the 2D and 3D condition did not reach statistical significance (p = 0.610).

Fig. 4 shows target-present and target-absent response times depending on the three test conditions.

![Fig. 4. Target-present and target-absent response times by condition (2D vs. 3D vs. OSARP). Error bars are ± one standard error.](image)

A mixed design ANOVA with condition (2D vs 3D vs OSARP) as between-subjects variable, trial type (target present vs target absent) as within-subjects variable and response time (in seconds) as the dependent variable showed a significant effect of condition, F(2, 59) = 48.32, p < 0.001, \( \eta^2 = 0.62 \), and an interaction between condition and trial type, F(2, 59) = 9.50, p < 0.001, \( \eta^2 = 0.24 \). Posthoc comparisons using Holm-Bonferroni corrections showed a significant difference between all three conditions (2D, 3D, OSARP) for target present trials (all p < 0.001) as well as for target-absent trials (all p < 0.001). Screeners needed more time with 3D imaging and OSARP than with 3D imaging without OSARP while 2D imaging resulted in the fastest response times. The comparisons between target-present and target-absent trials showed longer response times for target-present in the 2D condition (p = 0.022) and 3D condition (p < 0.001), but longer target-absent response times for the OSARP condition (p = 0.041).

These results are very interesting as both the 2D and 3D condition showed longer response times for target-present trials, while in the OSARP condition screeners needed more time for target-absent trials. Longer response times for the conditions using 3D imaging were expected as rotating and slicing obviously needs more time. However, to test whether the longer response times were also a result of the inexperience with the new 3D system, response times were compared between the two image blocks of the main test, to see whether screeners got faster over time. A mixed-design ANOVA with condition (3D vs. OSARP) as between-subjects variable, trial type (target-presents vs. target-absent) and block (1 vs 2) as within-subjects variables and response time as dependent variable showed a significant main effect of condition F(1, 41) = 29.04, p < 0.001, \( \eta^2 = 0.42 \), block F(1, 41) = 53.59, p < 0.001, \( \eta^2 = 0.57 \), and the interaction between condition and trial type, F(1, 41) = 14.28, p < 0.001, \( \eta^2 = 0.26 \). Post hoc comparisons were carried out for both conditions separately and showed significantly different response times between the first and second block for both target-absent and target-present trials (p < 0.001) for both the 3D imaging and OSARP condition with shorter response times for the second block.

### IV. SUMMARY, DISCUSSION AND CONCLUSION

In this study, we addressed two issues of theoretical and practical relevance. First, with the implementation of advanced 3D CT technology at airports currently using 2D multi-view X-ray systems, the question arises whether extensive and specific training is necessary for screeners who have never worked with 3D imaging before. Second, it was investigated whether a specific on-screen alarm resolution protocol (OSARP) would increase detection performance of screeners when using 3D imaging. We compared visual inspection performance of 2D hold baggage screeners by using three different test conditions: state-of-the-art 2D multi-view imaging versus newer 3D rotatable imaging versus 3D rotatable imaging following OSARP.

Screeners achieved similar detection performance with 2D compared to 3D imaging despite lower image quality of 3D imaging. The possibility of 360° rotation allowing visual inspection from all angles eliminates, or at least drastically reduces, the challenges resulting from low target visibility due to viewpoint or superposition effects in 2D imaging. This seems to compensate potential negative effects of lower image quality of 3D imaging. Moreover, our results suggest that screeners experienced with 2D multi-view X-ray imaging can transfer their visual inspection competency to 3D imaging. Considering this result, extensive training of 2D screeners before they can be allowed to work with 3D imaging is not needed.

Our study further showed that screeners’ detection performance could be increased when they followed a specific OSARP to visually inspect 3D rotatable images. This result is especially interesting as screeners only received a short 40 min training of the protocol. Due to the OSARP training, screeners also shifted their response bias to be more neutral (more biased toward judgement of target present). This implies that screeners were very compliant with the protocol resulting in more hits but also more false alarm decisions in the OSARP condition. However, the relevance of response bias results should be interpreted with caution, as they are dependent on target prevalence [24], [27]. With a significantly lower target prevalence in operation, a different response bias can be expected. Further, in operation, a longer training would be required for the use of such a protocol and therefore even larger improvements in detection performance might be expected.

Screeners needed more time when hold baggage was displayed with 3D imaging compared to 2D imaging and even more when OSARP was applied. It was anticipated that visual inspection with 3D imaging takes more time, because rotating and slicing 3D images obviously takes longer to process than a visual inspection of static 2D X-ray images. However, due to lower EDS alarm rates of 3D imaging systems compared to 2D
imaging systems in operation [2], [13], [15], [20], baggage throughput would still be substantially higher in hold baggage screening. Further, by instructing the screeners to explicitly follow the OSARP it is not surprising that screeners took more time. However, we found a significant decrease of response times in the second block of testing, which suggests that longer response times with OSARP decrease with practice.

As mentioned in the introduction, in HBS, the screeners’ task is mainly a decision task whereas in cabin baggage screening, visual inspection consists of search and decision making [3], [4]. Screeners in the 2D and 3D condition showed longer response times for target-present trials, whereas screeners in the OSARP condition needed more time for target-absent trials. A possible explanation is that in the 2D and 3D condition, screeners were able to quickly recognize EDS false alarms (i.e. when not all components of an IED are available), but took more time to confirm actual IEDs. This would also be consistent with the finding of a more conservative response bias.

Screeners in the OSARP condition took their decisions using a specific protocol following heuristic steps. They showed longer response times in target absent trials, which indicates that they took more time when an IED could not be identified. The more neutral response bias (more hits and more false alarms) indicates that the OSARP training could be improved further by adding heuristics to reduce false alarm decisions of screeners. However, it could also be possible that a longer training durations of the OSARP (compared to only 40 min like in this study) could already mitigate this problem. In summary, 2D screeners do not need extensive training to achieve comparable detection performance with 3D imaging. Training screeners with an OSARP increases detection performance but further research should be conducted to enhance OSARP training.

ACKNOWLEDGMENT
The authors particularly acknowledge the valuable contribution of Yotam Margalit during the whole process of the study. We also thank Melini Zeballos and Myra IsenSchmid for their help when conducting the data collection.

REFERENCES