The impact of color composition on X-ray image interpretation in aviation security screening
The impact of color composition on X-ray image interpretation in aviation security screening

Abstract

In order to improve aviation security, many airports apply Threat Image Projection (TIP) and computer-based X-ray image interpretation training (CBT). One difference between TIP and CBT X-ray images is the algorithm used to merge virtual threat items into X-ray images of passenger bags, resulting in different color nuances. In this study, we tested the influence of merging algorithms on threat object detection performance, reaction time and confidence rating of 12 airport security screeners. The image merging algorithms of the individually adaptive CBT X-Ray Tutor were used. We have found that the hit rate was higher for TIP images compared to CBT images. Accordingly, the mean of confidence ratings was increased for TIP images. The reaction times tended to be shorter for TIP images. The results of our study indicate that the CBT merging algorithm used in this study is more realistic than the tested TIP image merging algorithm.
The impact of color composition on X-ray image interpretation in aviation security screening

Claudia C. von Bastian  
Department of Psychology  
University of Zurich  
Zurich, Switzerland  
c.vonbastian@psychologie.uzh.ch

Adrian Schwaninger, Stefan Michel  
University of Applied Sciences Northwestern Switzerland  
School of Applied Psychology, Institute Human in  
Complex Systems  
4600 Olten, Switzerland  
and  
Center for Adaptive Security Research and Applications  
(CASRA)  
Thurgauerstr. 39  
8050 Zurich, Switzerland

Abstract—In order to improve aviation security, many airports apply Threat Image Projection (TIP) and computer-based X-ray image interpretation training (CBT). One difference between TIP and CBT X-ray images is the algorithm used to merge virtual threat items into X-ray images of passenger bags, resulting in different color nuances. In this study, we tested the influence of merging algorithms on threat object detection performance, reaction time and confidence rating of 12 airport security screeners. The image merging algorithms of the individually adaptive CBT X-Ray Tutor were used. We have found that the hit rate was higher for TIP images compared to CBT images. Accordingly, the mean of confidence ratings was increased for TIP images. The reaction times tended to be shorter for TIP images. The results of our study indicate that the CBT merging algorithm used in this study is more realistic than the tested TIP image merging algorithm.

Keywords-aviation security, display technologies, human-machine interaction, object recognition, image merging algorithms

I. INTRODUCTION

In aviation security, one main focus is the improvement of the process of X-ray screening of passenger bags in order to prevent forbidden objects getting past the security checkpoint. Although many airports are equipped with technologies of the newest generation, the detection of threat objects relies ultimately on human operators (airport security screeners), who visually inspect the X-ray images and decide whether a bag contains a threat object or not.

The performance of airport security screeners in the task of X-ray image interpretation is influenced by knowledge-based and image-based factors [1]; [2].

Knowledge-based factors refer to knowing which items are prohibited and what they look like in X-ray images of passenger bags. Some objects look quite different in X-ray images than in reality. Others, such as Improvised Explosive Devices (IED), are rarely seen in everyday life as well as at the security checkpoint and are therefore, difficult to recognize without the appropriate training.

According to [3], there are three image-based factors: rotation of the threat item, superposition by other objects, and bag complexity. Several studies have shown that the rotation of an object can have a strong impact on recognition (e.g., [4]; [5]; [6]; [7]). In general, X-ray images of forbidden objects are difficult to recognize when depicted from an unusual viewpoint and when diagnostic features are not visible. Another important factor contributing to image difficulty is the superposition of the threat object by other objects in a bag. For example, if a knife is superimposed by high density material, it becomes more difficult to recognize the characteristic shape of the object. Furthermore, the complexity of a bag, determined by the number and type of objects in the bag, has a significant influence on the detection performance.

Many airports approach the limiting factors described above with supportive measures like specific computer-based X-ray screening training (CBT) and Threat Image Projection (TIP).

It has been shown that CBT can substantially increase X-ray image interpretation competency and decrease reaction times [8]; [9]; [10]; [11]; [12]. Such training affects mainly knowledge-based factors and the detection of rotated objects [13]. Through training, airport’s security screeners learn which

Figure 1. Illustration of (a) a fictional threat item (FTI), (b) the position of a passenger bag where the FTI shall be inserted into, and (c) the resulting image which is shown to a security officer (screener) who decides whether the bag is OK or NOT OK.
objects are prohibited and what they look like in X-ray images. The screeners also store different and often unfamiliar views of the objects in visual memory [11].

TIP is a software function of current x-ray machines. Using this technology, fictional threat items (FTIs) are projected into X-ray images of real passenger bags during the routine airport security checks (see Fig. 1 for an illustration). The security officer ( screener) has to identify the potential threat and press a button. Screener responses are recorded and the system provides an immediate feedback. TIP has been developed with the aim to counter the typical human factor problems in traditional X-ray screening of passenger bags: low occurrence of threats, low interactivity of the screening task, and the difficulty of online performance measurement [14]. TIP became an important tool for enhancing the attention and vigilance of airport security screeners.

We have become aware that airport security screeners sometime state that they perceive TIP events as very easy due to non-authentic colors, unrealistic positions of the threat object within the bag, or an unrealistic threat item in a bag (e.g. a rifle in a small handbag), whereas they find CBT rather challenging. Consistent with these observations [15] found a ceiling effect, small inter-individual differences, and poor reliability for TIP data. TIP and CBT apply different algorithms for merging (or blending) FTIs into X-ray images of real passenger bags. Fig. 2 illustrates the different color nuances in an original X-ray image (Fig. 2a), a TIP image (Fig. 2b), and a CBT image (Fig. 2c).

Previous research has shown that in visual search tasks, a target in a unique color pops out from a display and guides the attention to this target [16]; [17]; [18]; [19]. Furthermore, [20] have shown that the larger the difference between target and distracters, the more efficient the search. This “pop-out” phenomenon has been attributed to a parallel-processing, pre-attentive mechanism that extracts unique perceptual features from cluttered visual scenes [21]; [22].

The goal of our study is to scrutinize whether the merging algorithm has an impact on the threat object detection performance of airport security screeners. We hypothesize that the color compositions created by the TIP merging algorithm lead to an increase of detection performance and confidence ratings for TIP images in comparison to CBT images, whereas the reaction times should be shorter for TIP images as for CBT images. Additionally, we expect reduced reaction times for the TIP images as a result of a pop-out effect due to unrealistic color compositions created by the TIP merging algorithm.

## II. Method

### A. Participants

We recruited 12 airport security screeners (3 males) with at least three years on-the-job experience at a big European airport. Screeners were tested previously with X-ray image interpretation test and achieved a high detection performance (i.e., $A' > .90$, and hit rate > .80) in the X-ray Competency Assessment Test (CAT, for details see [23]), as well as in the Bomb Detection Test (BDT, for details see [24]). Additionally, the recruited airport security screeners had to meet our criteria of a high on-the-job performance in 2007 recorded by the TIP system (i.e., $A' > .90$, hit rate > .85, and false-alarm rate < .15). $A'$ is a “nonparametric” detection performance measure, which takes the hit rate as well as the false-alarm rate into account [25]; [26].

The age of the participants ranged from 26 to 55 (M = 45.00 years, SD = 8.79 years). They were all naïve with regard to the hypotheses under investigation.

### B. Material

The performance test consisted of X-ray images of bags, which were cabin baggage captured with X-ray machines at a European airport using the auto-archive function. These images were revised by three airport security supervisors in order to remove inappropriate images (e.g., images containing prohibited items or liquids). We used 512 of these bags with a high bag complexity, calculated by using the formula for opacity displayed in (1) as described by [27].

\[
OP = \frac{\sum_{x,y} I_N(x,y) < 64}{BS}
\]

The formula reflects the extent to which X-rays are able to penetrate objects in a bag. $I_N(x,y)$ denotes the pixel intensities, whereas 64 is the pixel intensity threshold beneath which the pixels are counted. $BS$ is the size of the bag and is used to standardize the Opacity value on bag size.

Each bag was used twice, once combined with a prohibited item, and once without any threat object. The threat objects belong to four categories of prohibited items: guns, knives,
Improvised Explosive Devices (IED), and other prohibited items (e.g., gas, chemicals, or grenades). The threat objects have been captured by experts of Zurich State Police, Airport division.

We used 16 exemplars of each category and combined every item with two different bags, in a manner that the degree of superposition by other objects was high. For this purpose we used the formula for superposition depicted in (2) [27].

$$SP = C - \sum_{x,y} (I_{SN}(x,y) - I_{N}(x,y))^2$$

The function computes the difference between the pixel intensity values of the bag image with the threat object ($I_{SN}(x,y)$) and the pixel intensities of the corresponding harmless bag ($I_{N}(x,y)$).

Furthermore, each threat item was presented in two different rotations. The easy rotation shows the object from a canonical perspective [28] as judged by two security experts who captured the stimuli. The difficult rotation shows the threat item rotated horizontally or vertically by 85 degrees relative to the canonical view.

Each of these combined threat images was once created by using the TIP merging algorithm, and once by using the CBT merging algorithm. Hence, the images were identical, with the exception that we varied the merging algorithm (see Fig. 3).

Overall, the experiment comprised 1024 trials: 16 (threat objects) * 4 (categories) * 2 (rotation) * 2 (bag with/without threat item) * 2 (each threat object combined with two bags) * 2 (CBT or TIP merging algorithm).

C. Procedure

After two practice trials, half of the participants started with the CBT condition, the other half with the TIP condition. Each trial was presented for 15 seconds. Participants had to decide whether the presented bag was OK (contains no threat item) or NOT OK (contains a threat item) by clicking the respective button on the screen. Participants were instructed to answer as fast and as accurately as possible. Additionally, they were asked to indicate how confident they were in their decision by clicking on a slider on the screen. The 1024 trials have been subdivided into four blocks. Participants were allowed to take a short break after completing each block. Trials were randomized within each block. Completing the experiment took about 90 minutes.

III. Results

A. Analysis of hit and false-alarm rates

In order to examine the effect of the merging algorithm on the detection performance, we analyzed hit and false-alarm rates. The hit rate refers to the proportion of all images containing a prohibited item that have been judged as NOT OK, while the false-alarm rate refers to the proportion of NOT OK judgments for harmless bags. There was a significant increase of the hit rate for the TIP merging algorithm in comparison to the CBT merging algorithm, $t(11) = -29.65 , p < .001$, with a large effect size of $d = 1.67$, while the false-alarm rate remained the same, $t(11) = -0.19, p = .43, d = 0.02$ (see Fig. 4).

![Figure 3. (a) A CBT trial containing a gun in an easy rotation and (b) the corresponding TIP trial.](image)

![Figure 4. Effect of the merging algorithm on the hit and false-alarm rate.](image)

Note that absolute performance values are not reported due to security reasons.

Next, we ran a two-way analysis of variance (ANOVA) for repeated measures using the hit rate with the within-participant factors merging algorithm (TIP, CBT) and category (guns, knives, IED, and other threat objects) in order to compare the
effect of the merging algorithm on the detection performance regarding the different threat object categories. We found a large main effect of the merging algorithm, \( F(1, 11) = 878.83, p < .001, \eta^2 = .99 \), as well as of the category, \( F(3, 33) = 231.61, p < .001, \eta^2 = .96 \). The interaction between merging algorithm and category was also significant, \( F(3, 33) = 10.30, p < .001, \), with a large effect size of \( \eta^2 = .48 \). Pairwise comparisons revealed a significant increase of detection performance for guns, \( t(11) = -19.08, p < .001, d = 1.60, \) knives, \( t(11) = -15.91, p < .001, d = 1.71, \) IED, \( t(11) = -13.12, p < .001, d = 1.59, \) and other threat objects, \( t(11) = -20.64, p < .001, d = 1.60. \) According to [29], all these effect sizes are large (see Fig. 5).

According to [29], all these effect sizes are large (see Fig. 5).

![Figure 5](image)

**Figure 5.** Effect of the merging algorithm on the threat object category. Note that absolute performance values are not reported due to security reasons.

![Figure 6](image)

**Figure 6.** Effect of the merging algorithm on the reaction time pooled across all categories and for each category. Error bars represent standard errors of the mean (SEM).

### B. Analysis of reaction times

The effect of the merging algorithm on the reaction time pooled across all categories was marginally significant, \( t(11) = 1.74, p = .06 \), with a medium effect size \( d = 0.56 \) [29]. Fig. 6 shows the tendency of a prolonged reaction time for threat images produced with the CBT merging algorithm. A two-way ANOVA for repeated measures using the reaction time with the within-participant factors merging algorithm (TIP, CBT) and category (guns, knives, IED, and other threat objects) revealed no effect of the merging algorithm, \( F(1, 11) = 3.02, p = .11, \eta^2 = .22 \), but for category, Greenhouse-Geisser \( F(1.43, 15.68) = 23.52, p < .001, \eta^2 = .68 \). The interaction between merging algorithm and category was also not significant, Greenhouse-Geisser \( F(1.59, 17.40) = 2.26, p = .14, \eta^2 = .17 \). Pairwise comparisons disclosed that the difference in the reaction time was only significant for the category IED, \( t(11) = -13.12, p < .01, d = 0.98 \). For all other categories, the merging algorithm had no significant influence (guns, \( t(11) = 0.81, p = .22, d = 0.25 \), knives, \( t(11) = 1.49, p = .08, d = 0.52 \), and other threat objects, \( t(11) = 1.29, p = .11, d = 0.47 \). Although the effects did not meet the level of significance, the effect sizes are at least small (guns and other threat objects) or medium (knives) according to [29].

### C. Analysis of confidence ratings

Last, we analyzed the confidence ratings. Pooled across all categories, the mean of confidence ratings was higher for TIP images, \( t(11) = -3.28, p < .01, d = 0.48 \) (see Fig. 7). A two-way ANOVA for repeated measures using the mean of confidence ratings with the within-participant factors merging algorithm (TIP, CBT) and category (guns, knives, IED, and other threat objects) revealed a large main effect of the merging algorithm, \( F(1, 11) = 10.73, p < .01, \eta^2 = .49 \), as well as of the category, Greenhouse-Geisser \( F(1.28, 14.04) = 4.67, p < .05, \eta^2 = .30 \). The interaction between merging algorithm and category was also significant, \( F(3, 33) = 4.80, p < .01, \eta^2 = .30 \). Pairwise comparisons disclosed a significant increase of perceived confidence for guns, \( t(11) = -1.88, p < .05, d = 0.32, \) knives, \( t(11) = -4.61, p < .001, d = 0.66, \) IED, \( t(11) = -3.57, p < .01, d = 0.49, \) and other threat objects, \( t(11) = -2.54, p < .05, d = 0.40. \)

![Figure 7](image)

**Figure 7.** Effect of the merging algorithm on the confidence ratings pooled across all categories and for each category. Error bars represent standard errors of the mean (SEM).

### IV. Discussion

The goal of our study was to examine whether there are differences between merging algorithms of TIP and CBT which would impact detection performance, reaction time and experienced confidence in X-ray image interpretation.

Consistent with our first hypothesis, it was revealed that the hit rate is higher when CTIs were created using the TIP merging algorithm compared to CTIs created with the CBT merging algorithm. This effect was found for all threat object categories (guns, knives, IED, and other prohibited items).
Consistent with [20] and our second hypothesis, reaction times were longer for CBT images in comparison to TIP images (although this effect was only marginally significant). The effect meets the level of significance for the category IED with a large effect size, whereas the effect sizes for the other categories are, apart from not being significant, at least small or medium. A larger sample size might lead to significant differences.

The increase of experienced confidence for TIP CTI confirms our third hypothesis pooled across all conditions as well as for each category.

These findings are consistent with the results of [15], which showed a ceiling-effect, small inter-individual differences and poor reliability of TIP data. Additionally, the reduced reaction times for TIP images indicate that the increased detection performance might in fact be mediated through unique color artifacts created by the TIP merging algorithm.

Overall, our study provides evidence that the merging algorithm influences the detection performance of professional airport security screeners. Furthermore, we demonstrated that the CBT merging algorithm used in this study is more realistic than the TIP merging algorithm and therefore lays the better foundation for an effective training of airport security screeners.

ACKNOWLEDGMENT

We are very thankful to Zurich State Police, Airport Division for their help in creating the stimuli and the good collaboration for conducting the study. We also thank Anton Bolfing for his valuable help with the image analyses.

REFERENCES


