TOWARDS A MODEL FOR ESTIMATING IMAGE DIFFICULTY IN X-RAY SCREENING

A. Schwaninger, S. Michel, A. Bolfing

Department of Psychology, University of Zurich, Switzerland

ABSTRACT

In this study we developed a first computational model for estimating image difficulty of x-ray images of passenger bags. Based on [1] three image-based factors are proposed as predictors of image difficulty: view difficulty of the threat item, superposition by other objects, and bag complexity (i.e. clutter and transparency of the bag). First, these factors were validated using detection experiments. We then developed computer-based algorithms to estimate the image-based factors automatically. Finally, we could show that our computational model can better explain human performance than human ratings of the image-based factors.

1. INTRODUCTION

The relevance of aviation security has increased dramatically in the last years. One of the most important tasks is the visual inspection of passenger bags using x-ray machines. In this study we investigated the role of image-based factors on human detection of prohibited items in x-ray images. Schwaninger has proposed in [1] that the following image-based factors influence how difficult it is to detect a threat item in x-ray images: view difficulty of the threat item, superposition by other objects, and bag complexity. This was validated in a study conducted recently [2]. In Experiment 1, we replicated these results in order to provide converging evidence for the validity of the assumption of different image-based factors. In Experiment 2, the same x-ray images were rated by human participants for view difficulty, superposition, bag complexity (clutter and transparency), and general difficulty. These human ratings were then correlated with detection performance obtained in the first experiment. In Experiment 3, we developed computer-based algorithms to estimate the image-based factors automatically. These estimates were correlated with human ratings of the same image-based factors (obtained in Experiment 2). Using multiple linear regression analysis, we examined in Experiment 4 whether our computer-based estimates were able to predict human performance from Experiment 1 as good as human ratings from Experiment 2 on the same image-based factors could do so.

In this study, only results from guns are presented. A more detailed study containing data from guns and knives is going to be published later.

2. EXPERIMENT 1

The main aim of Experiment 1 was to replicate the results of [2], in which it was shown that view difficulty, superposition, and bag complexity influence detection performance substantially.

2.1. Method and Procedure

2.1.1 Participants

Twelve undergraduates of the University of Zurich participated in this study (5 males, 7 females). None of them had participated in a study with x-ray images before.

2.1.2 Procedure

The Object Recognition Test (ORT) was used to analyze the influence of the three image-based factors view difficulty, superposition and bag complexity on human detection performance (for details see [2] and [3]). X-ray images of passenger bags were shown 4 seconds each. Participants had to decide whether a bag is OK (no threat item present) or NOT OK (threat item present). Using a slider control, participants indicated on a 90 point rating scale how sure they were in their decision (confidence ratings). There were a total of 256 test trials: 16 (8 guns, 8 knives) x 2 (easy vs. difficult view) x 2 (low vs. high superposition) x 2 (low vs. high bag complexity) x 2 (threat bag vs. harmless bag). No feedback was given on test trials. Prior to the test trials, 8 practice trials were presented followed by a presentation of the threat items. The 8 guns were shown for 10 seconds followed by a 10 second screen with the 8 knives. Half of the items were shown in easy view, the other half in difficult view (for further details see [2] and [3]).

2.1.3 Statistical Analysis

In the study conducted by [2] detection was measured using A' (for details on this and other detection measures see [4] and [5]). In this study, we were interested in developing a computational model to explain detection performance of threats in x-ray images. In Experiment 1 we calculated hit rates for each participant by averaging across threat images. Individual hit rates were subjected to a three-way analysis of variance (ANOVA) with view difficulty, superposition and bag complexity as within-participant factors.

2.2 Results and Discussion

The main effects are illustrated in Figure 1. All were highly significant with large effect sizes (η^2 values). View difficulty: $\eta^2 = .95$, $F(1,11) = 211.2$, $p < .001$; superposition: $\eta^2 = .49$, $F\ (1,11) = 10.5$, $p < .01$; bag complexity: $\eta^2 = .59$, $F(1,11) = 16.0$, $p < .01$. This replicates earlier findings in which large main effects of view, superposition and bag complexity were found for A' scores [2].

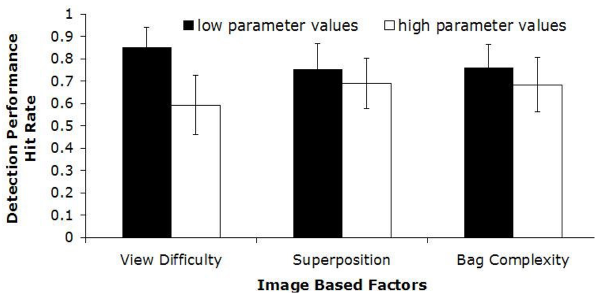

Figure 1. Illustration of main effects of view difficulty, superposition and bag complexity on hit rates for guns.

Only one significant interaction was found: Bag complexity * view difficulty: $\eta^2 = .77$, $F(1,11) = 36.4$, $p < .001$. All other interactions were not significant. This is consistent with the assumption of three relatively independent factors (whereas only view difficulty and bag complexity might interact).

3 EXPERIMENT 2

3.1 Introduction

The main aim of Experiment 2 was to investigate whether human ratings of view difficulty, superposition and bag complexity are correlated with human performance measured in Experiment 1.

3.2 Method and Procedure

3.2.1 Participants

The same participants of Experiment 1 took part in Experiment 2 (with a delay of about one week).

3.2.2 Procedure

The same x-ray images as in Experiment 1 were used. The participant's task was to rate view difficulty and superposition of the threat items (threat bags only), and clutter, transparency and general image difficulty (threat and non-threat bags). Ratings were given using a graphical slider control (from very low = 0 to very high = 50). Prior to the ratings, 8 practice trials were presented.

3.2.3. Statistical Analysis

Hit rates per x-ray image were calculated by averaging performance data from Experiment 1 across participants. These hit rates were then correlated with x-ray image ratings on threat bags from Experiment 2 (per image, averaged across participants).

3.3 Results and Discussion

Pearson correlations showed that ratings of view difficulty and superposition were significantly correlated with hit rate, $r(64) = -.521$, $p < .001$, and $r(64) = -.522$, $p < .001$, respectively. The other correlations did not reach statistical significance: Hit rate and clutter, $r(64) = -.17$, $p = .19$; hit rate and transparency, $r(64) = .08$, $p = .56$. These results could suggest that both clutter and transparency are not relevant for the detection of the threat items used in this study (only guns, see introduction), or that the participants could not reliably estimate the degree of clutter and transparency. We are currently conducting further research to investigate these possibilities.

4 EXPERIMENT 3

4.1 Introduction

In Experiment 3, computer-based estimates for image-based factors were developed. They were compared to human ratings from Experiment 2 in order to determine their perceptual plausibility. The following table shows the abbreviations for all independent variables.

Independent Variables	Computer-based Estimates	Rating Estimates
View Difficulty	VD_C	VD_R
Superposition	SP_C	SP_R
Clutter	CL_C	CL_R
Transparency	TR_C	TR_R

Table 1. Abbreviations used in this article. Indices C and R represent computer-based and human rating estimates, respectively.

4.2 Method and Procedure

4.2.1 Computer based estimates

Computer-based estimates were developed for view difficulty, superposition, and bag complexity (i.e. clutter and transparency).

4.2.1.1 View Difficulty

View difficulty VD was calculated by averaging hit rates ($pHit_i$) across different threat images displaying the same threat item view. In the ORT, each threat item is displayed 4 times from the same viewpoint (see section 2.1.2.). The detection performance of the item in question ($pHit_j$) was excluded from this average detection performance. This was done in order to avoid a circular argument in the statistical model by partial inclusion of a predictor into the criterion variable (see section 5). Therefore, the n in the view difficulty formula equals 4, but the average was calculated over the remaining three (n-1) images displaying the same threat item view.

Formula 1 $$\textit{View Difficulty}\quad VD_j = \frac{\left(\left(\sum_{i=1}^{n} pHit_i\right) - pHit_j\right)}{n-1}$$

4.2.1.2 Superposition

The computer-based estimate of superposition is based on the Euclidian distance between the grayscale pixel intensities of the bag with the threat item (I_{SN}) and the bag without it (I_N). The following formula was used:

$$\textit{Superposition} \quad SP = \sqrt{\sum (I_{SN}(x,y) - I_N(x,y))^2}$$

Formula 2

4.2.1.3 Clutter

Clutter (CL) should represent the amount of disarrangement in the bag. In our approach it was estimated based on the amount of high pass frequency information:

$$\textit{Clutter}\, CL = \sum (I_N(x,y) \otimes F^{-1}(HP(f_x, f_y)))$$

$$HP(f_x, f_y) = 1 - \frac{1}{1 + \left(\frac{\sqrt{(f_x^2 + f_y^2)}}{f}\right)^d}$$

Formula 3

This convolution ($\otimes$) is equivalent to a high-pass Butterworth filter application in the Fourier-space (F^{-1}: inverse Fourier transform), where f_x and f_y are the frequency components, f is the cut-off frequency and d the fall off.

4.2.1.4 Transparency

Metallic content is more difficult to penetrate by x-ray than organic material, which therefore appears more "transparent" or less opaque in the x-ray image. Transparency was estimated based on the number of pixels in the darkest quarter (< 65) of the pixel intensity range (0 to 255), relative to the bags overall size (areas with pixel intensities $\neq$ 255).

$$\textit{Transparency} \quad TR = \frac{\sum_{x,y}(I_N(x,y) < 65)}{\sum_{x,y}(I_N(x,y) \neq 255)}$$

Formula 4

4.2.2 Statistical Analysis

To examine the perceptual plausibility of the computer-based estimates we calculated their correlations with the corresponding human ratings from Experiment 2.

4.3 Results and Discussion

As can be seen on the diagonal of the correlation matrix of Table 2, all correlations between computer-based estimates and human ratings were highly significant (except for clutter). This shows that at least three of the four of our computer-based estimates of image-based factors are perceptually plausible. The high correlation between computer-based estimates of transparency and human ratings for clutter could indicate that our participants had problems in distinguishing between clutter and transparency. This is consistent with the high correlation between human ratings of clutter and transparency, r(64) = -.79, p < .001.

	VD_R	SP_R	CL_R	TR_R
VD_C	-.61**	-.32**	-.06	-.00
SP_C	-.22	-.44**	-.28*	.15
CL_C	-.04	.12	.15	-.10
TR_C	-.03	.32**	.67**	-.62**

*p<.05. **p<.01

Table 2. Correlations between computer-based estimates and human ratings.

5 EXPERIMENT 4

5.1 Introduction

The aim of this experiment was to examine how well our computer-based estimates can explain human performance.

5.2 Method and Procedure

Multiple linear regression analysis was used to test how well our computer-based estimates of image-based factors can explain human performance measured in Experiment 1. The human ratings from Experiment 2 were used for benchmarking. More specifically, we tested whether our computer-based estimates of image-based factors achieve a better prediction of human performance than human ratings of the same image-based factors.

5.2.1 Statistical Analysis

The two equations below show the two multiple linear regression models using computer-based estimates (C indices) and human ratings (R indices) of image-based factors. The abbreviation DP represents detection performance (hit rate per image averaged across participants), which is the dependent variable. The two models were compared in terms of their goodness-of-fit measures, their regression coefficient's significances, and – most importantly – the percentage of variance in the dependent variable the models were able to explain.

$$DP = b_0 + b_1 VD_C + b_2 SP_C + b_3 CL_C + b_4 TR_C + R$$
$$DP = b_0 + b_1 VD_R + b_2 SP_R + b_3 CL_R + b_4 TR_R + R$$

5.3 Results and Discussion

Note that the scales of the computer-based estimates and the rated image based factors have opposite signs. Therefore, the beta-weights in predicting the dependent variable (hit rate per image) have opposite signs in the computational and the rating models.

5.3.1 Computational Model

The computational model correlates with human performance with r = .76 (Figure 2). As shown at the bottom of Table 3, our model using computer-based estimates is able to explain 55 % of the variance of the hit rate (adjusted R^2).

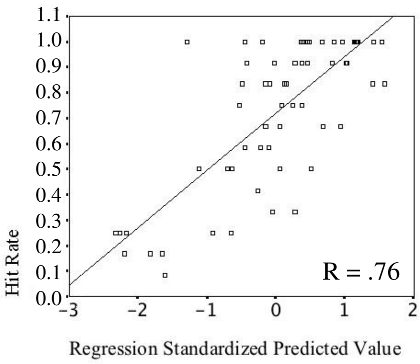

Figure 2. Correlation between predicted and observed performance using the computational model.

Interestingly, view difficulty and superposition explain most of the variance of the hit rate. In fact, only their beta weights are significant (Table 3).

Variable	B	SE B	β
VD_C	0.78	0.10	.68**
SP_C	0.02	0.01	.23*
CL_C	-0.00	0.00	-.01
TR_C	-0.08	0.45	-.02

R^2 = .581, R^2(adj) = .553, $F(4,59) = 20.455$, $p < 0.001$
*p < .05. **p <.01.

Table 3. Summary of regression analysis using computer-based estimates of image-based factors for predicting hit rates.

5.3.2 Human Ratings Model

The model based on human ratings correlates with human performance with r = .70 (Figure 3).

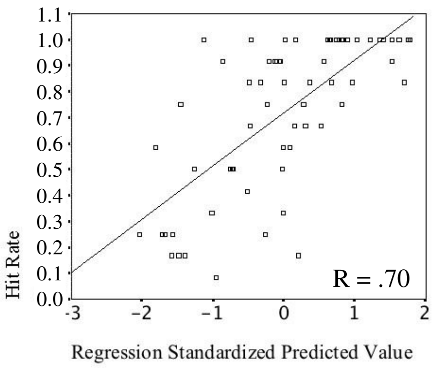

Figure 3. Correlation between predicted and observed performance using the human ratings model.

As can be seen at the bottom of Table 4, the human ratings were able to explain 45 % of the variance of the hit rate (adjusted R^2). This means, that our computational model could explain human performance better than a model based on human ratings. Interestingly, for both models, view difficulty and superposition explained most of the variance of the hit rate and the beta weights of clutter and transparency were not significant.

Variable	B	SE B	β
VD_R	-0.01	0.00	-.46**
SP_R	-0.02	0.00	-.48**
CL_R	-0.00	0.00	-.05
TR_R	-0.01	0.01	-.11

R^2 = .485, R^2(adj) = .452, $F(4,59) = 14.004$, $p < 0.001$
*p < .05. **p < .01.

Table 4. Summary of regression analysis using human ratings of image-based factors for predicting hit rates.

6 GENERAL DISCUSSION

This study provided converging evidence for the view that detection performance in x-ray screening depends on view difficulty, superposition and bag complexity [1]. The results of Experiment 1 showed large main effects of these image-based factors on human detection performance, which is highly consistent with earlier findings [2]. Human ratings (Experiment 2) and computer-based estimates (Experiment 3) were significantly correlated for view difficulty and superposition. Using multiple regression it was shown in Experiment 4 that our computational model could explain human performance (hit rate) better than a model based on human ratings. Interestingly, for both models, view difficulty and superposition explained most of the variance of the hit rate. In contrast, bag complexity (clutter and transparency) was a weak predictor for both, the model based on human ratings, as well as the computational model. As explained in the introduction, only results from guns are presented in this study. We are currently conducting a series of experiments using different threat types and computer-based estimates in order to extend the computational model presented in this paper and to further investigate the role of bag complexity.

7 ACKNOWLEDGMENT

We are thankful to Zurich State Police, Airport Divison and Zurich Airport Unique for supporting this study. Thanks to Franziska Hofer for her help in preparing the manuscript.

8 REFERENCES

[1] Schwaninger, A. (2003). Evaluation and selection of airport security screeners. *AIRPORT, 02/2003*, 14-15.
[2] Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38,* 258-264.
[3] Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). The X-Ray Object Recognition Test (X-Ray ORT) – a reliable tool for measuring visual abilities needed in x-ray screening, *this volume*.
[4] Green, D.M. & Sweets, J.A. (1996). *Signal detection theory and psychophysics,* New York: Wiley.
[5] Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38,* 303-308.