

A statistical approach for automated image difficulty estimation in x-ray screening using image processing algorithms

Anton Bolfig

University of Zurich, Department of Psychology, Visual Cognition Research Group (VICOREG),
Switzerland

Stefan Michel and Adrian Schwaninger

University of Zurich, Department of Psychology, Visual Cognition Research Group (VICOREG),
Switzerland

Abstract

In this study we present a model for automated image difficulty estimation in x-ray screening. Schwaninger, Hardmeier, and Hofer (2004, 2005) have identified three image based factors that affect x-ray image difficulty: view difficulty depending on the rotation of an object, superposition by other objects, and bag complexity depending on clutter and transparency of objects. Experiment 1 is a replication of earlier findings confirming the effects of these image based factors on human detection performance. In Experiment 2 significant correlations between human ratings of the image based factors and human detection performance were found. In Experiment 3 image processing algorithms were developed to automatically estimate x-ray image difficulty based on view difficulty, superposition and bag complexity. Significant correlations with human detection performance and human ratings of the image-based factors indicate high perceptual plausibility of the computer estimates. In Experiment 4 we apply multiple linear regression analysis to compare a linear model based on computer estimates with a linear model based on human ratings of the image-based factors. It was shown that our computer model can predict human performance at least as good as human ratings. Applications for threat image projection and adaptive computer-based training systems are discussed.

Introduction

The aim of this study is to evaluate our statistical model for image difficulty assessment. Threat image projection (TIP) is a software function available on modern x-ray equipment that allows inserting fictional threat items (FTIs) into x-ray images of real passenger bags. TIP is a source of motivation to screeners, provides a means of improving screeners' threat knowledge, and can be used to assess screeners' threat detection performance. Schwaninger, Hardmeier and Hofer (2004) have identified three major image-based factors influencing detection performance: View difficulty of the FTI depending on its rotation, superposition by other objects in the bag, and bag complexity, i.e. clutter, the texture's unsteadiness, and transparency, the relative size of dark areas. Current TIP systems project FTIs into real passenger bags based on a random ratio into a random position of the bag. As a consequence, TIP images vary substantially in image difficulty depending on effects of view difficulty, superposition and bag complexity. When TIP data is used to assess screener performance such effects make it difficult to obtain reliable measurements. The main aim of the current work is to develop a computational model using image processing in order to determine x-ray image difficulty while taking effects of view difficulty, superposition, and bag complexity into account.

The study comprises four experiments. The first experiment is a replication of earlier findings to confirm the relevance of the three major image-based factors in predicting human detection performance. In the second experiment we estimated the relevance of these image-based factors by correlating subjective ratings of them with the hit rate ($p(\text{hit})$) in detection performance. We expect high correlations to reflect high influence of the subjectively perceived image-based factors on the measured item difficulty $p(\text{hit})$. In Experiment 3 we correlated the computer-based estimates with human ratings to estimate the perceptual plausibility of our computer algorithms for image-based factors estimation. Additionally, this allows us to

The 4th International
Aviation Security Technology Symposium
November 27 - December 1, 2006

check for possible intercorrelations among the predictors, which allows us to detect statistical dependencies among them. Finally, in experiment 4 we used multiple linear regression analyses to compare our computational model and human perception with regard to how well they can predict human detection performance.

Experiment I: ORT Findings Replication

Method and Procedure

Participants

The sample size of participants was twelve undergraduate students in psychology (5 females). None of the participants has had experience with x-ray images before.

ORT Test Design

Stimuli were displayed on 17" TFT screens at a distance of about 100 cm so that x-ray images subtended approximately 10-12 deg of visual angle. The computer program measured outcome (hit, miss, false alarm (FA), correct rejection (CR)) and the time from image onset to final decision key press.

In this study we used the X-Ray Object Recognition Test (Hardmeier, Hofer, und Schwaninger, 2005) which contains 256 x-ray images, half of them with an FTI (threat images), the other half without an FTI, i.e. non-threat images. Viewpoint difficulty, superposition and bag complexity are counterbalanced using the following design: 16 (threat items, i.e. 8 guns and 8 knives) x 2 (easy vs. difficult viewpoint) x 2 (easy vs. difficult superposition) x 2 (easy vs. difficult bag complexity) x 2 (threat vs. non-threat images).

Procedure

The X-Ray ORT is fully computer-based. Before starting the test, several practice trials are presented to make sure that the task is understood properly. The participant's task is to decide whether a bag is OK (no threat item present) or NOT OK (threat item present). Each x-ray image disappears after 4 seconds. No feedback is given to their answers. In this study, only trials containing knives have been used for analysis because of their high familiarity.

Statistical Analysis

A three-way analysis of variance (ANOVA) with view difficulty, superposition, and bag complexity as within-participant factors was used on percentage of detected threats (hit rate) per participant and factor combination.

Results

The following results were obtained in the ANOVA. There were clear main effects of view difficulty: $\eta^2=.84$, $F(1,11)=59.06$, $p<.001$, and superposition: $\eta^2=.65$, $F(1,11)=20.48$, $p<.001$. The effect of bag complexity was only marginally significant: $\eta^2=.23$, $F(1,11)=3.30$, $p=.10$. Interaction effects: View difficulty * bag complexity: $\eta^2=.52$, $F(1,11)=11.93$, $p<.01$ and view difficulty * superposition * bag complexity: $\eta^2=.34$, $F(1,11)=5.60$, $p<.05$.

Discussion

We found clear effects of view difficulty and superposition in hit rates, which replicates the results of an earlier study using the X-Ray ORT with novices and screeners by calculating A' scores (Hardmeier, Hofer, and Schwaninger, 2005; Schwaninger, Hardmeier et al., 2005). Thus, the results of Experiment 1 provide further evidence for the validity of image-based factors as important determinants of x-ray image difficulty. However, the effect of bag complexity was only marginally significant whereas in earlier studies large main effects of bag complexity were found.

Experiment II: Image-based Factors Rating

Method and Procedure

Rating

The same experimental setup was used as in Experiment 1. The participant's task in Experiment 2 was to rate the X-Ray ORT images in terms of general difficulty and the following image-based factors: View difficulty, superposition, and bag complexity (clutter and transparency). Each of these dimensions could be rated on a continuing scale between "very low" to "very high" using a slider control.

**The 4th International
Aviation Security Technology Symposium**
November 27 - December 1, 2006

Statistical Analysis

In order to validate the influence of the subjectively perceived image-based factors on the hit rate we correlated the image-based factors ratings with p(hit) derived from Experiment 1 using Pearson's product-moment correlation.

Results

The correlation between objective hit rate and subjectively rated view difficulty was $r(64)=-.53$, $p<.001$, between p(hit) and superposition $r(64)=-.67$, $p<.001$, between p(hit) and clutter $r(64)=-.24$, $p=.06$, and between p(hit) and transparency $r(64)=.31$, $p<.05$.

Discussion

The results show impressively that there is a close contiguity between objective hit rates and the subjective perception of our image-based factors view difficulty, superposition, and bag complexity. This indicates a high perceptual plausibility of these image based factors. However, bag complexity, comprising clutter and transparency, shows the lowest correlation, whereas clutter does not correlate significantly with the hit rate.

An explanation could be that bag complexity is difficult to rate because of low perceptual plausibility. Indeed, the two components of bag complexity (clutter and transparency) were highly correlated $r(64)=-.89$, $p<.001$. This could imply that novices have a hard time in distinguishing the two components of bag complexity and just give similar ratings for both.

Experiment III: Image-based Factors

Introduction

Experiment 3 was designed to develop image-processing algorithms for estimating view difficulty, superposition, and bag complexity automatically in x-ray images. These algorithms were then validated by correlating the computer-based estimates with the human ratings from Experiment 2.

Method and Procedure

Image Processing Algorithms

All image-processing algorithms developed for this purpose are based on theoretical considerations. For each image-based factor the consequences of high and low parameter values of each single image-based factor on the pixel and frequency space have been determined. Different algorithm parameters were optimized by maximizing the correlations between the image-based factor estimates and human detection performance of earlier studies (Hardmeier et al., 2005; Schwaninger, Hardmeier et al., 2005; Schwaninger, Michel, and Bolting, 2005).

In the following subsections the resulting image-processing algorithms are described separately.

View Difficulty

Because it is not possible to determine the degree of 3-D rotation (view difficulty) of a physical threat item solely from the 2-D x-ray image, this image-based factor is not being calculated by image-processing, but statistically from X-Ray ORT detection performance data.

The image-based factor view difficulty is basically calculated by averaging the hit rates of a certain threat item in one of the two views presented in the ORT. In order to avoid a circular argument in the statistical model (multiple linear regression) by partial inclusion of a predictor into the criterion variable, the detection performance of the one item in question is being excluded from this average detection performance estimate.

$$VD_j = \frac{\left(\left(\sum_{i=1}^n p(\text{hit})_i \right) - p(\text{hit})_j \right)}{n-1}$$

In this study, each threat item (knives only) is being displayed four times in the same view, 2 (bag complexity low vs. high) x 2 (superposition low vs. high). Therefore, the n in the view difficulty formula equals 4, but the average is calculated over the remaining three items.

Superposition

The image-processing algorithm for superposition simply calculates the Euclidian distance between the grayscale pixel intensities of the signal-plus-noise (SN) image and the harmless bag (N, noise) image.

$SP = \sqrt{\sum_{x,y} (I_{SN}(x,y) - I_N(x,y))^2}$ For each pixel of a threat image, the pixel intensity difference between the bag with the threat item and the bag without the threat item is calculated and squared. All squared pixel intensity differences are summed up. The final image-based factor is then derived from calculating the square root of this sum of squared pixel intensity differences.

Clutter

This image-based factor is designed to express bag item properties like their texture unsteadiness, disarrangement, chaos or just clutter. In terms of the depicted bags themselves, this factor is closely related to the amount of items in the bag as well as their structures in terms of complexity and fineness. The method used in this study is based on the assumption, that such texture unsteadiness can be described mathematically in terms of the amount of edges, i.e. the amount of transitions in luminosity within a certain space frequency range surpassing a certain threshold.

$CL = \sum_{x,y} (I_N(x,y) \otimes F^{-1}(HP(f_x, f_y)))$ We implemented this mathematical formulation by first applying on the intensity image of the empty bag a convolution kernel, which is derived from a highpass-filter in the Fourier space by inverse Fourier transformation. In a second step, the amount of the resulting pixels, representing edges as described above are being counted.

Transparency

The image-based factor transparency reflects the amount to which x-rays are able to penetrate various objects in the bag. This depends on the specific material density of these objects. Heavy metallic materials such as lead are known to be very hard to be penetrated by x-rays. For a screener, the consequence is that he cannot see any objects in front or on the back of such material.

$TR = \frac{\sum_{x,y} (I_N(x,y) < thresh)}{\sum_{x,y} (I_N(x,y) \neq 255)}$ The implementation of the image-processing algorithm for the image-based factor transparency consists in the calculation of the amount of pixels being darker than a certain threshold (<thresh) of the pixel intensity range going from 0 to 255, relative to the bags overall size (< 255, white pixels).

Correlations

To evaluate the perceptual plausibility of these image-processing algorithms for estimating view difficulty, superposition, and bag complexity we correlated them with the human ratings obtained in Experiment 2.

Results

The correlations were $r(64)=-.47, p<.001$ for view difficulty, $r(64)=-.44, p<.001$ for superposition, $r(64)=.18, p=.16$ for clutter and $r(64)=-.63, p<.001$ for transparency.

Discussion

Except for clutter all correlations between calculations and ratings are highly significant. Remember the high intercorrelation between the human ratings of the image-based factors clutter and transparency ($r(64)=-.89, p<.001$) obtained in Experiment 2. Here, in Experiment 3 we found a quite high intercorrelation between the corresponding calculated image-based factors clutter and transparency ($r(64)=-.55, p<.001$). Therefore, we must keep in mind that these two factors are not fully independent. This is not very surprising as we subsume them within the factor bag complexity. Nevertheless we can conclude that our calculations are compatible with the perceptual plausibility of novice screeners.

Experiment IV: Statistical Model Introduction

Experiment 4 was designed to evaluate the predictive power of our computational model and to compare it to human perception as a tough baseline.

Method and Procedure

One multiple regression analysis was carried out using the computationally estimated image-based factors, and the other using the subjectively rated image-based factors as predictors.

Multiple Regression Analysis

To achieve the goals of our statistical model we can use multiple linear regression analysis whereas our image-based factors are the predictors, and the hit rate of human observers obtained in Experiment 1 is the dependent variable.

In the following, we describe the computational model and the model based on human ratings of perceived view difficulty, superposition and bag complexity.

Linear model using computationally calculated image-based factors as predictors:

$$DP = b_0 + b_1VD_c + b_2SP_c + b_3CL_c + b_4TR_c + R$$

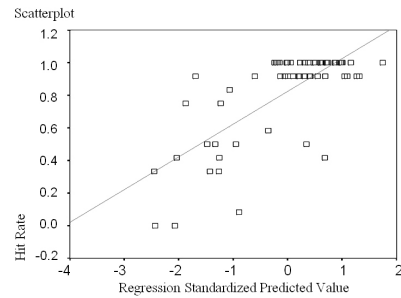
Linear model using mean values of the subjectively rated image-based factors as predictors:

$$DP = b_0 + b_1VD_r + b_2SP_r + b_3CL_r + b_4TR_r + R$$

Results

Computational Model

The figure to the right shows the scatterplot of the multiple linear regression analysis using the calculated image-based factors as predictors with the standardized predicted values on the abscissa and the measured human performance (hit rates) on the ordinate.



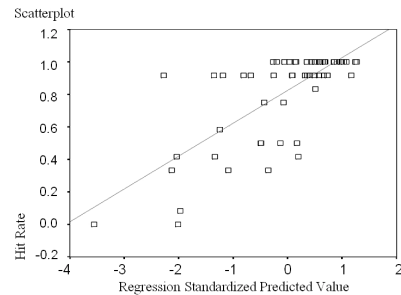
In the table below the basic statistical criteria of the multiple linear regression using calculated image-based factors are listed:

Variable	B	β
VD _c	.352	.288*
SP _c	.056	.497***
CL _c	.000	-.149
TR _c	-.172	-.029

$R^2=.543$, $R^2(adj)=.512$, $F(4,59)=17.49$, $p<0.001$
 $p<.05$, ** $p<.01$, *** $p<.001$

Ratings Model

The figure to the right shows the scatterplot of the multiple linear regression analysis using the subjectively rated image-based factors as predictors with the standardized predicted values on the abscissa and the measured ORT hit rates on the ordinate.



In the table below the basic statistical criteria of the multiple linear regression using subjectively rated image-based factors are listed:

Variable	B	β
VD _r	-.010	-.329**
SP _r	-.021	-.675**
CL _r	-.005	-.131
TR _r	-.013	-.338

$R^2=.548$, $R^2(adj)=.518$, $F(4,59)=17.91$, $p<0.001$
* $p<.05$, ** $p<.01$.

Discussion

In Experiment 4 we have developed a computational model to calculate view difficulty, superposition, and bag complexity automatically in x-ray images using image processing algorithms. These algorithms were significantly correlated with human ratings of view difficulty, superposition, and bag complexity. This result shows that our image processing algorithms are perceptually plausible. In order to benchmark our model we compared its predictive power with a model based on human perception. Our computational model using image processing algorithms for automatic estimation of view difficulty, superposition, and bag complexity could predict human detection performance in the X-Ray ORT with a correlation between model prediction and measured performance of $r = .74$. In order to benchmark our model we have compared it to a linear model using human ratings of view difficulty, superposition, and bag complexity. As we can see from the R^2 values of the regression analyses our computational model performed equally well as a linear model using ratings of perceived view difficulty, superposition, and bag complexity. It is important to focus on the strength of the impacts of the single image-based factors used as predictors, the beta-weights. As we expected based on Experiment 2 and Experiment 3, the image-based factors solely depending on the empty bags contribute little to the prediction of the hit rate. In both, the computational model as well as the ratings model, only view difficulty and superposition result in significant beta-weights.

General Discussion

The results show clearly, that it is possible to develop an automatic system that calculates x-ray image difficulty using image processing algorithms. Further research needs to be done regarding the division of image-based factors predicting hit and false alarm rates. The non-threat image based predictor bag complexity may rather predict false alarm rates than hit rates. Furthermore, there are great chances to enhance the existing predictors and possibly add some undiscovered ones as well. Especially in the field of bag complexity measures, probably the most challenging ones, there is quite some work to be done in the future. As mentioned earlier in this paper, this study was conducted with knives as threat items only. Schwaninger, Michel, et al. (2005) have shown comparable effects for guns based on the same study design as presented here: Guns calculations model with $R^2(\text{adj})=.578^{***}$, β -weights .614^{***} for view difficulty and .327^{***} for superposition and the ratings model with $R^2(\text{adj})=.595^{***}$, β -weights .388^{***} for view difficulty and .602^{***} for superposition. Other threat categories like improvised explosive devices (IEDs) and others are expected to result in different priorities of view difficulty, superposition, and bag complexity. Such studies can be highly valuable to find out how different properties of threats in terms of their shape and material can influence human detection performance and visual search.

Acknowledgment

We are thankful to Zurich State Police, Airport Division for their help in creating the stimuli and the good collaboration for conducting the study.

References

- Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). The object recognition test (ORT) – a reliable tool for measuring visual abilities needed in x-ray screening. *IEEE ICCST Proceedings*, 39, 189-192.
- Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings*, 38, 303-308.
- Schwaninger, A., "Object recognition and signal detection," in *Praxisfelder der Wahrnehmungspsychologie*, B. Kersten and M.T. Groner, Eds., Bern: Huber, in press.
- Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. *IEEE Aerospace and Electronic Systems*, 20(6), 29-35.
- Schwaninger, A., Michel, S., & Bolting A. (2005). Towards a model for estimating image difficulty in x-ray screening. *IEEE ICCST Proceedings*, 39, 185-188.