A Statistical Approach for Image Difficulty Estimation in X-Ray Screening Using Image Measurements

Adrian Schwaninger*
Department of Psychology,
University of Zurich, Switzerland
Max Planck Institute for Biological Cybernetics,
Tübingen, Germany

Stefan Michel†
Department of Psychology,
University of Zurich, Switzerland
Max Planck Institute for Biological Cybernetics,
Tübingen, Germany

Anton Bolfing‡
Department of Psychology,
University of Zurich, Switzerland
Max Planck Institute for Biological Cybernetics,
Tübingen, Germany

Abstract

The relevance of aviation security has increased dramatically at the beginning of this century. 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, Hardmeier, and Hofer (2004, 2005) have identified three image based factors: View Difficulty, Superposition and Bag Complexity. This article consists of 4 experiments which lead to the development of a statistical model that is able to predict image difficulty based on these image based factors. Experiment 1 is a replication of earlier findings confirming the relevance of image based factors as defined by Schwaninger et al. (2005) on x-ray detection performance. In Experiment 2, we found significant correlations between human ratings of image based factors and human detection performance. In Experiment 3, we introduced our image measurements and found significant correlations between them and human detection performance. Moreover, significant correlations were found between our image measurements and corresponding human ratings, indicating high perceptual plausibility. In Experiment 4, it was shown using multiple linear regression analysis that our image measurements can predict human performance as well as human ratings can. Applications of a computational model for threat image projection systems and for adaptive computer-based training are discussed.

Keywords: Airport security, image difficulty estimation, image measurements, image metrics, statistical modeling, x-ray screening

1 Introduction

The relevance of aviation security has increased dramatically in recent years and there has been substantial progress regarding screening technology, especially in the field of automatic explosive detection systems [Ying et al. 2006]. However, the last decision is always made by a human operator and investigating human factors as essential determinants of security screening performance has become an important research topic. First contributions in the field of x-ray image inspection were based on research in medical image interpretation [Gale et al. 2000]. Krupinski et al. (2003) were able to identify important factors that influence pulmonary nodule detection. Experimental psychology studies [Ghylin et al. 2006] and eye movement research [McCarley et al. 2004; Liu et al. 2006] have been useful to better understand visual search and perceptual learning in x-ray image interpretation. A series of studies conducted in recent years has provided converging evidence for the importance of scientifically based selection, training, and testing methods to achieve and maintain high levels of performance in x-ray image interpretation [Schwaninger 2005b; Schwaninger 2006b].

The aim of this study is to develop and evaluate a statistical model for image difficulty estimation in x-ray screening using image measurements. Schwaninger et al. (2005) could show that there are three major image based factors which affect detection performance: View difficulty depending on the rotation of an object, superposition by other objects in the bag, and bag complexity, which comprises clutter, the bag's background texture unsteadiness, and transparency, the relative size of dark areas in the bag. Figure 1 illustrates the three image based factors as proposed by Schwaninger, Hardmeier and Hofer (2005).

A model for image difficulty estimation using automated image measurements and human performance statistics can be very useful for threat image projection (TIP) data analysis and adaptive computer based training (CBT). TIP is a software function of state-of-the art x-ray machines which allows the automated insertion of fictional threat items (FTIs) into x-ray images of real passenger bags. TIP systems are operational in several countries and used to enhance motivation and attention of screeners on the job. Since the TIP to bag ratio is relatively low (i.e. the number of projections per passenger bags) and the resulting TIP images (x-ray image of real passenger bag plus FTI) vary substantially with regard to image based factors, it is difficult to obtain reliable individual performance measurements. With a reliable statistical model for image difficulty estimation using image measurements, corrected individual performance scores could be calculated, which would allow more reliable individual performance assessments. A second application is adaptive CBT. For example, the individually adaptive algorithms of X-Ray Tutor start with easy views of threat items shown in bags of low complexity with little superposition by other objects. Once a threat item is recognized by a screener, the view difficulty is increased and it is shown in more complex bags with more superposition (for details on X-Ray Tutor see [Schwaninger 2004b]). There are large differences between individuals regarding their ability to

*e-mail: a.schwaninger@psychologie.uzh.ch
†e-mail: s.michel@psychologie.uzh.ch
‡e-mail: a.bolfing@psychologie.uzh.ch

APGV 2007, Tübingen, Germany, July 26–27, 2007.
© 2007 ACM 978-1-59593-670-7/07/0007 $5.00

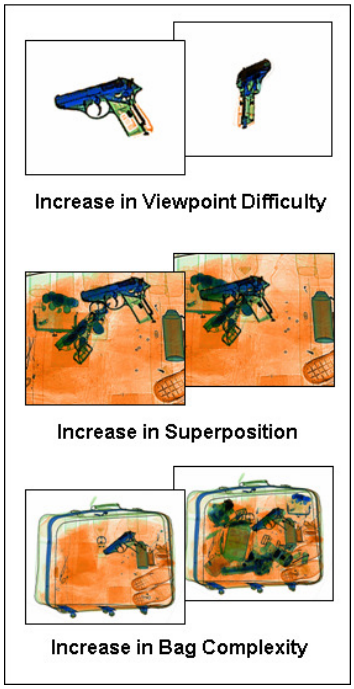

Figure 1: *Illustration of the three major image based factors suggested by Schwaninger, Hardmeier and Hofer (2005).*

cope with image-based factors (Schwaninger et al., 2005). Therefore, a good model for image difficulty estimation using automated image measurements of image-based factors could be very useful for enhancing such individually adaptive training algorithms.

The study is sectioned into four experiments. The first experiment is a replication of earlier findings [Schwaninger et al. 2005a] to confirm the relevance and relative independence of image based factors in predicting human performance. The second experiment aims to estimate the subjective perceptual plausibility of the underlying image based factors by correlating them with the average hit rate ($p(hit)$), i.e. percent detection per image averaged across participants. Threat images were rated for view difficulty, superposition, clutter, transparency and general difficulty. Images of harmless bags were rated for clutter, transparency, and general difficulty. The correlation between these ratings and human detection performance reflects the relative importance of each image based factor. We then developed statistical formulae and automated image measurements for the above mentioned image based factors. Experiment 3 was designed to estimate the perceptual plausibility of these computer generated estimates. We correlated the computer-based estimates with the corresponding human ratings to determine whether our computer-based algorithms correspond with human perception. Finally, in Experiment 4 we compared a model using computer-based estimates to a model based on human ratings of the image based factors.

2 Experiment 1

2.1 Method

Experiment 1 is a replication of the study by Schwaninger et al. (2005), who identified image based factors for threat item detection in x-ray image screening. Two important differences need to be mentioned. In view of possible applications in TIP systems, we are mainly interested in predicting the percentage of correct responses to images containing a threat item. Therefore, we use the hit rate instead of d' as the variable to be predicted. In our previous studies, we used the signal detection measure $d' = z(H) - z(FA)$ whereas $z(H)$ refers to the z-transformed hit rate and $z(FA)$ to the z-transformed false alarm rate [Green and Swets 1966]. Secondly, only novices and no experts are tested because we want to examine image based factors independent of expertise.

2.1.1 Participants

Twelve undergraduate students in psychology from the University of Zurich participated in this experiment (5 females). None of them has had any previous experience with visual inspection of x-ray images.

2.1.2 Materials

The X-Ray Object Recognition Test (X-Ray ORT) was used to measure detection performance. This test has been designed to analyze the influence of image based effects view difficulty, superposition and bag complexity on human detection performance when visually inspecting x-ray images of passenger bags. Inspired by signal detection theory [Green and Swets 1966], the X-Ray ORT consists of two sets of 128 x-ray images. One set contains harmless bags without a threat item (N-trials, for noise). The other set contains the same bags, each of them with a threat (SN-trials, for signal-plus-noise). Only guns and knives of typical familiar shapes are used. This is important because the X-Ray ORT is designed to measure cognitive visual abilities to cope with effects of viewpoint, superposition, and bag complexity independent of specific visual knowledge about threat objects. The X-Ray ORT consists of 256 items (x-ray images) given by the following test design: 16 threat item exemplars (8 guns, 8 knives) x 2 view difficulty levels x 2 bag complexity levels x 2 superposition levels x 2 trial types (SN and N-trials). The construction of the items in all image based factor combinations as shown above was lead by visual plausibility criteria. After choosing two sets of x-ray images of harmless bags with different parameter values in bag complexity, the sixteen fictional threat items were projected into the bags in two different view difficulties at two locations with different superposition each. The term fictional threat items (FTIs) is commonly used in connection with TIP systems as discussed in the introduction. For further details on the X-Ray ORT see [Hardmeier et al. 2005; Schwaninger et al. 2005a]. Stimuli were displayed on 17" TFT screens at a distance of about 100cm, so that the x-ray images subtended approximately 10-12 degrees of visual angle. The computer program measured outcome (hit, miss, false alarm, correct rejection) and the response times from image onset to final decision button press.

2.1.3 Procedure

X-ray images of passenger bags were shown for a maximum display duration of 4 seconds. Note that at airport security controls the human operators (screeners) usually have only 3-6 seconds to inspect a passenger bag. The participant's task was to decide whether the image is OK (i.e. the bag contains no threat item) or NOT OK (i.e. it contains a threat item) by clicking one of the corresponding buttons on the screen (see Figure 2). In addition, participants had to judge their confidence using a slider control (from UNSURE to SURE). These confidence ratings were used for another study. No feedback was given regarding the correctness of the responses. Participants could initiate the next trial by pressing the space bar.

Several practice trials were presented to make sure that the task was understood properly before the test started. Immediately prior to the actual test, all guns and knives were presented on the screen for 10

Figure 2: *Screenshot of an X-Ray ORT trial showing an x-ray image of a passenger bag containing a gun. Response buttons and slider control are aligned at the bottom of the screen.*

seconds, respectively. This was done to minimize any effects of threat item knowledge. Half of the items were shown in easy view and the other half in difficult view.

2.2 Results

2.2.1 Descriptive Results

Figure 3 displays the mean hit rate ($M = .80$) and standard deviation ($SD = 0.17$) broken up by main effects of view difficulty, superposition, and bag complexity for guns and knives. Data was first averaged across images for each participant and then across participants to calculate mean hit rate. The analysis of false alarm rates is not part of this study but will eventually be published later.

2.2.2 Statistical Analyses

Our hypothesis whereby the image based factors have great influence on detection performance was tested using repeated-measures ANOVA. Main effects are stated below.

Guns:

View Difficulty:	$\eta^2 = .89$	$F(1,11) = 91.55$	$p < .001$
Superposition:	$\eta^2 = .40$	$F(1,11) = 7.45$	$p < .05$
Bag Complexity:	$\eta^2 = .14$	$F(1,11) = 1.76$	$p = .21$

Knives:

View Difficulty:	$\eta^2 = .84$	$F(1,11) = 59.06$	$p < .001$
Superposition:	$\eta^2 = .65$	$F(1,11) = 20.48$	$p < .001$
Bag Complexity:	$\eta^2 = .23$	$F(1,11) = 5.60$	$p = .10$

2.3 Discussion

We were able to replicate the results from Schwaninger et al. (2005) involving professional screeners fairly well regarding main effects of view difficulty and superposition. However, unlike in earlier studies, the image based factor bag complexity had no significant effect on the hit rate for both, guns and knives. The most probable reason for this is that the detection performance in this study is the hit rate instead of d'. As mentioned earlier, d' equals $z(H)-z(FA)$ whereas H refers to hit rate and FA to false alarm rate (Green & Swets, 1966). Effects of bag complexity are more likely to be found on false alarm rate. In x-ray screening tests, the false alarm rate is based on the number of times a participant judges a bag to be NOT OK even though there is no threat item in it. Consistent with this view, we found clear effects of bag complexity on d' in earlier studies [Hardmeier et al. 2005; Schwaninger et al. 2005a]. It is therefore not so surprising that we could not find a significant effect of bag complexity on hit rate alone in Experiment 1.

3 Experiment 2

Experiment 2 was designed to investigate the perceptual plausibility of our image measurements introduced in section 4.

3.1 Method

The participants who had conducted Experiment 1 took part in Experiment 2 one week later using a slightly modified experimental setup. The participant's task was to rate the difficulties of the X-Ray ORT images regarding view difficulty and superposition of the threat images. In addition, clutter, transparency and general item difficulty had to be rated for threat and non-threat images. The ratings were given by mouse clicks on a 50-point scale (0=very low to 50=very high). No initial position was set. Figure 4 shows a screenshot.

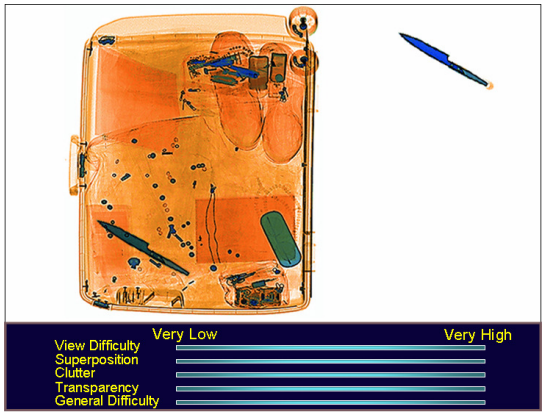

Figure 4: *Screenshot of a typical trial of Experiment 2 containing a knife. All participants were asked to judge the image based factors subjectively, whereby bag complexity is separated in clutter and transparency. Additionally, participants were asked to judge the general item difficulty as well (not analyzed in this study). Threat items were displayed next to the bag. For non-threat items, the slider controls for view difficulty and superposition were discarded.*

3.2 Results

In order to estimate the relative importance of image based factors [Schwaninger et al. 2005a] on human detection performance, we correlated ratings for view difficulty, superposition, clutter and transparency (Experiment 2) with the hit rate data obtained in Experiment 1. Data analysis was conducted separately for guns and knives.

3.2.1 Descriptive Results

Figure 5 shows the averaged ratings across all participants and across all threat items. The ordinate depicts the rating scores on the 50-point scale (see Figure 4). The black and white bars in each image based factors category represent the low and high parameter values according to the arrangement of the X-Ray ORT test design. Over-all mean rating value was $M = 19.2$ with a standard deviation of $SD = 15.4$. Inter-rater consistency was quite high with an average correlation (Fisher-corrected) between subjects of $r = .64$ for view difficulty, $r = .62$ for superposition, $r = .65$ for clutter and $r = .40$ for transparency.

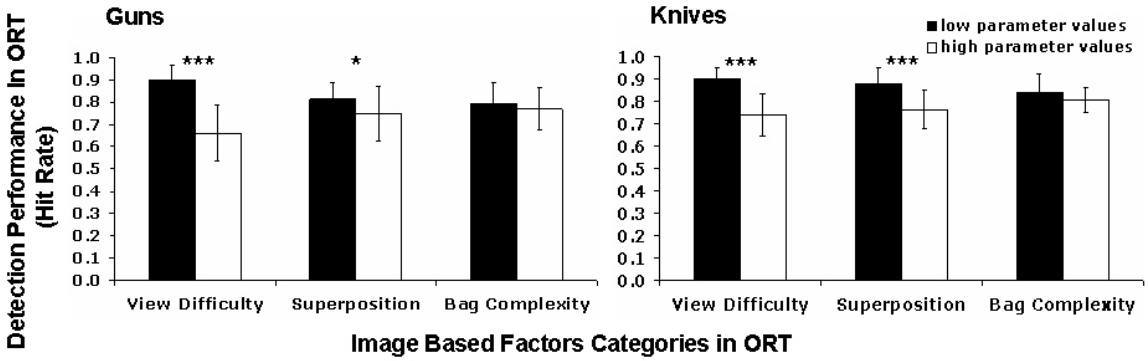

Figure 3: *Results of Experiment 1. Mean hit rate for the detection of guns and knives, broken up by main effects of view difficulty, superposition, and bag complexity. Data was first averaged across images for each participant and then across participants to calculate mean hit rate. Error bars represent the standard deviation across participants.*

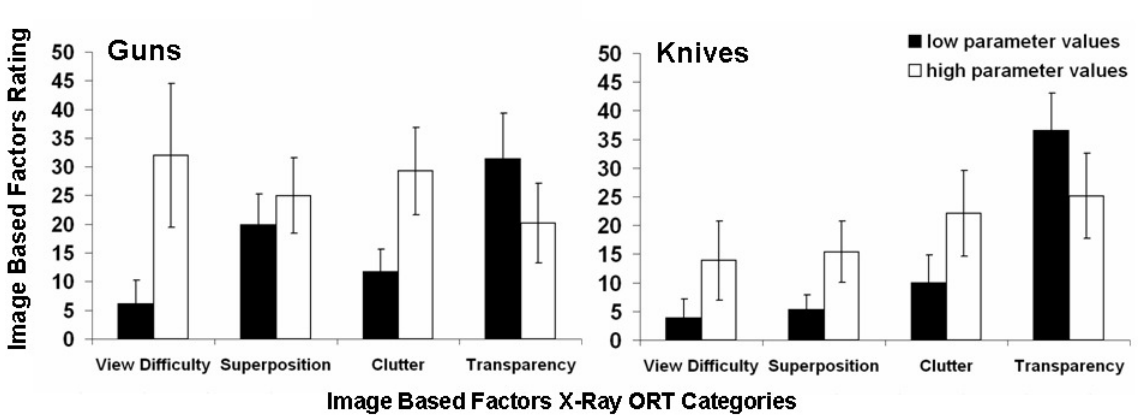

Figure 5: *Descriptive results from Experiment 2 for guns and knives separately. The image based factor bag complexity from the X-Ray ORT is split into the sub-factors clutter and transparency according to the rating experiment design shown in Figure 4. Please note that the factor transparency points in the opposite direction compared to bag complexity and the other image based factors.*

3.2.2 Statistical Analyses

Correlations of ratings of image based factors with hit rate per image averaged across the participants of Experiment 1.

Guns:

View Difficulty:	$r(12) = -.56$	$p < .001$
Superposition:	$r(12) = -.69$	$p < .001$
Clutter:	$r(12) = -.32$	$p < .05$
Transparency:	$r(12) = .37$	$p < .01$

Knives:

View Difficulty:	$r(12) = -.53$	$p < .001$
Superposition:	$r(12) = -.67$	$p < .001$
Clutter:	$r(12) = -.24$	$p = .06$
Transparency:	$r(12) = .31$	$p < .05$

Concerning the mathematical signs, note that the hit rate points in the opposite direction of threat detection difficulty. The more difficult a threat item is to be detected the lower the hit rate.

3.3 Discussion

All subjective human ratings show significant correlations with the hit rates from Experiment 1, except for clutter in x-ray images containing a knife, which was marginally significant ($p = .06$). Thus, the results from Experiment 1 and Experiment 2 showed that image based factors affect objective x-ray image difficulty (hit rate) and the image-based factors can be rated by novices. Consistent with the findings from Experiment 1, the ratings of image based factors show that clutter and transparency are less predictive than ratings of view difficulty and superposition. For the development of image measurements, it was necessary to split up the factor bag complexity into clutter and transparency. However, this seems to be problematic, because for subjective ratings they seem to be highly interdependent. The ratings of clutter and transparency are highly correlated: $r(12) = -.93, p < .001$ for guns and $r(12) = -.86, p < .001$ for knives. We return to this issue in section 4.3.

4 Experiment 3

The aim of Experiment 3 was to develop computer-based algorithms to automatically estimate the image based factors view difficulty, superposition, clutter, and transparency. The perceptual plausibility of these computer-based algorithms was examined by correlating them with the human ratings obtained in Experiment 2.

4.1 Method

All image measurements developed for this purpose are based on theoretical considerations. Different algorithm parameters were optimized by maximizing the correlations between the image-based factors estimates and detection performance measures derived from earlier X-Ray ORT findings from x-ray screening experts.

4.1.1 Statistical estimates and image measurements for image based factors

View Difficulty
Even with the aid of 3D volumetric models, it is not (yet) possible to satisfyingly determine the degree of a 3-dimensional rotation (view difficulty) of a physical threat item automatically from its 2-dimensional x-ray image [Mahfouz et al. 2005]. Additional difficulties regarding image segmentation arise from the very heterogeneous backgrounds of x-ray images, compare [Sluser and Paranjape 1999]. Therefore, this image based factor is not (yet) being calculated by image processing, but statistically from X-Ray ORT detection performance data obtained in Experiment 1.

$$VD_j = \frac{\left(\sum_{i=1}^{4} \text{HitR}_i\right) - \text{HitR}_j}{3} \tag{1}$$

Equation 1 shows the calculation of the image based factor view difficulty, whereas i is the summation index ranging from 1 to 4 (2 bag complexities x 2 superpositions), j denotes the index number of the x-ray image in question (one threat exemplar in one of the two views), $HitR_j$ is its average hit rate across all participants and '4' is the number of the bags each FTI was projected into. In order to avoid a circular argument in the statistical model (multiple linear regression, see Experiment 4) by partial inclusion of the criterion variable into a predictor, the hit rate of the one item in question is excluded from this estimate.

It is important to understand that this concept of view difficulty is not just reflecting the degree of rotation of an object. In that case there would be two parameter values for all threat exemplars only. View difficulty as it is conceptualized here reflects innate view difficulty attributes unique to each exemplar view separately.

Superposition
This image based factor refers to how much the pixel intensities at the location of the FTI in the threat bag image differ from the pixel intensities at the same location in the same bag without the FTI. Equation 2 shows the image measurement formula for superposition. $I_{SN}(x, y)$ denotes the pixel intensities of a threat image and $I_N(x, y)$ denotes the pixel intensities of the corresponding harmless bag.

$$SP = \sqrt{\sum_{x,y} \left(I_{SN}(x, y) - I_N(x, y)\right)^2} \tag{2}$$

It should be noted that this mathematical definition of superposition is dependent on the size of the threat item in the bag. For further development of the computational model it is conceivable to split up superposition and the size of the threat item into two separate image based factors. Measurement of superposition would require having both the bag with the FTI and without. For both applications mentioned in the introduction, this is possible with current TIP and CBT technology. In TIP, the FTI, its location, the bag with and without the FTI are recorded. In several CBT systems, the same information is recorded and stored, too.

Clutter
This image based factor is designed to express bag item properties like its textural unsteadiness, disarrangement, chaos or just clutter. In terms of the bag images presented, this factor is closely related to the amount of items in the bag as well as to 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 high frequency regions.

$$CL = \sum_{x,y} I_{hp}(x, y) \tag{3}$$
$$\text{where} \quad I_{hp}(x, y) = I_N * \mathcal{F}^{-1}(hp(f_x, f_y))$$

Equation 3 shows the image measurement formula for clutter. It represents a convolution of the empty bag image (N for noise) with the convolution kernel derived from a high-pass filter in the Fourier space. I_N denotes the pixel intensities of the harmless bag image. $\mathcal{F}^{-1}$ denotes the inverse Fourier transformation. $hp(f_x, f_y)$ represents a high-pass filter in the Fourier space (see Appendix).

Transparency
The image based factor transparency reflects the extent to which x-rays are able to penetrate objects in a bag. This depends on the specific material density of these objects. These attributes are represented in x-ray images as different degrees of luminosity. Heavy metallic materials such as lead are known to be very hard to be penetrated by x-rays and therefore appear as dark areas on the x-ray images.

$$TR = \frac{\sum_{x,y}(I_N(x, y) < \text{threshold})}{\sum_{x,y}(I_N(x, y) < 255)} \tag{4}$$

Equation 4 shows the image measurement formula for transparency. $I_N(x, y)$ denotes the pixel intensities of the harmless bag. $threshold$ is the pixel intensity threshold beneath which the pixels are counted. The implementation of the image measurement for the image based factor transparency is simply achieved by counting the number of pixels being darker than a certain threshold (e.g. < 65) relative to the bag's overall size (< 255, non-white pixels).

4.2 Results

To examine perceptual plausibility of the computer-based measurements, we correlated them with the corresponding averaged ratings from Experiment 2.

4.2.1 Statistical Analyses

Pearson's product-moment correlations between the calculated measurements and the corresponding human ratings' mean values were applied for each image based factor dimension separately.

Guns:

View Difficulty:	$r(12) = -.62$	$p < .001$
Superposition:	$r(12) = -.54$	$p < .001$
Clutter:	$r(12) = .16$	$p = .20$
Transparency:	$r(12) = -.69$	$p < .001$

Knives:

View Difficulty:	$r(12) = -.47$	$p < .001$
Superposition:	$r(12) = -.44$	$p < .001$
Clutter:	$r(12) = .18$	$p = .16$
Transparency:	$r(12) = -.63$	$p < .001$

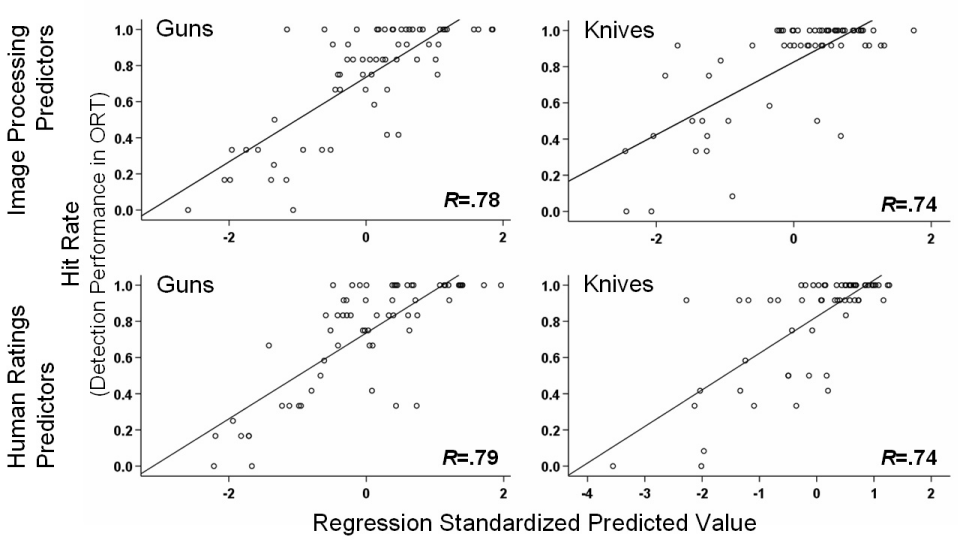

Figure 6: *The four scatter plots from the models predicting the hit rate on the basis of all disposable image based factors as predictors. Guns and knives are displayed separately (columns). The models based on the calculated predictors derived from image measurements are displayed in the first row and the models based on rated image based factors predictors are displayed in the second row.*

4.3 Discussion

Except for clutter all correlations between automated measurements and ratings are highly significant. In the discussion of Experiment 2 the high inter-correlations between the human ratings of the image based factors clutter and transparency was mentioned ($r(12) = -.93, p < .001$ for guns and $r(12) = -.86, p < .001$ for knives). Consistent with this result, there were also fairly high inter-correlations between the corresponding calculated estimates of the image based factors clutter and transparency ($r(64) = .52, p < .001$ for guns and $r(64) = .55, p < .001$ for knives). Except for clutter, we can conclude that our algorithms for automated estimation of image based factors are perceptually plausible because they correlate significantly with the ratings of novices.

5 Experiment 4

Experiment 4 was designed to evaluate the predictive power of a statistical model based on automated estimation of image based factors. To this end, we now compare the results of multiple linear regression analysis using the automated estimates of image based factors as predictors with the results of multiple linear regression analysis using the human ratings of image based factors as predictors.

5.1 Method

The comparison included the four image based factors introduced in Experiment 3.

5.1.1 Multiple Linear Regression Analysis

The predictors of the multiple linear regression model are our image based factors; the hit rates per image averaged across participants (Experiment 1) is the dependent variable. We compared the two statistical models in terms of their goodness-of-fit measures, their regression coefficient's significances and the percentage of variance in the dependent variable hit rate the model is able to explain by its predictors.

5.2 Results

5.2.1 Descriptive Results

Figure 6 shows the scatter plots with regression standardized predicted values on the abscissa and the actually measured hit rate from Experiment 1 on the ordinate.

5.2.2 Statistical Analyses

Figure 7 shows the most important statistical values of the four multiple linear regression analyses arranged in columns and rows like in Figure 6. The single tables show the four predictors in the rows. The first column gives the variable names of the image based factors. Unstandardized regression weights are in the second and standardized beta weights in the third column. Column four shows the p-value statistics indicating the significance of the single regression coefficients in the model. The last line shows the goodness-of-fit statistics of the model as a whole. R^2 tells us to which extent the model is able to predict the variance in the hit rate. Because R^2 increases with the number of predictors independently of their predictive power, $R^2(adj)$ taking into account the number of predictors is given, too. Finally the statistical indices F-value and the significance level of the model as a whole (p-value) are given.

All statistical models are highly significant in the overall goodness-of-fit verification statistics, both for guns and knives. The R^2-values, the extent to which a model is able to explain the variance in the dependent variable by its predictors, are very high compared to values usually obtained when predicting human performance. The model based on our image measurements achieves an R^2 of .61 ($R^2(adj)$=.58) with guns and an R^2 of .54 ($R^2(adj)$=.51) with knives. The ratings model is even marginally better with an R^2 of .62 ($R^2(adj)$=.60) with guns and an R^2 of .55 ($R^2(adj)$=.52) with knives. Concerning the regression coefficients in detail, the predictors view difficulty and superposition are always significant, mostly highly significant. This is not the case for the two sub-factors of bag complexity (clutter and transparency).

Image measurements

Guns

Variable	Regression weight	Beta-weight β	Significance p-value
View Difficulty	0.72	.614	.000
Superposition	0.023	.327	.000
Clutter	0.000033	.114	.239
Transparency	-0.82	-.176	.072

R^2=.605, R^2(adj)=.578, F(4,59)=22.60, p<.001

Knives

Variable	Regression weight	Beta-weight β	Significance p-value
View Difficulty	0.35	.288	.018
Superposition	0.057	.497	.000
Clutter	-0.00003	-.149	.194
Transparency	-0.17	-.029	.788

R^2=.543, R^2(adj)=.512, F(4,59)=17.49, p<.001

Human ratings

Variable	Regression weight	Beta-weight β	Significance p-value
View Difficulty	-0.009	-.388	.000
Superposition	-0.018	-.602	.000
Clutter	0.01	.355	.227
Transparency	0.011	.273	.107

R^2=.621, R^2(adj)=.595, F(4,59)=24.15, p<.001

Variable	Regression weight	Beta-weight β	Significance p-value
View Difficulty	-0-01	-.329	.001
Superposition	-.0.21	-.675	.000
Clutter	-0.003	-.131	.497
Transparency	-0.17	-.338	.123

R^2=.548, R^2(adj)=.518, F(4,59)=17.913, p<.001

Figure 7: *Statistical analysis tables of the models with the most important statistical values of the multiple linear regression analyses. Each of the four tables shows the statistical values of the verification of each regression coefficient separately in the rows. Additionally the model's overall goodness-of-fit verification values are given in the bottom row of each model. In both statistical models, the dependent variable is the hit rate obtained in Experiment 1.*

5.3 Discussion

The different statistical models in Experiment 4 show that the image based factors suggested by Schwaninger et al. (2005) are quite powerful predictors of human detection performance. The model based on automated estimation of image-based factors is as predictive as human ratings. Admittedly, Experiment 4 shows also that the subfactors of the image based factor bag complexity, clutter and transparency, do not contribute significantly to the explanatory power of the model. In some cases, they even show regression weights which point in the direction opposite to what is expected. As already mentioned in Experiments 1 and 2 this can be explained by the fact that in detection experiments bag complexity rather affects the false alarm rate than the hit rate. This is currently being investigated in additional experiments. Nevertheless, our computational model is able to predict the hit rate in terms of image based factors as good as human ratings can. Such a model could therefore provide a basis for the enhancement of individually adaptive computer-based testing and training systems in which the estimation of x-ray image difficulty is an essential component. In addition, the image measurements developed in this study can be very useful for analyzing more reliable individual TIP performance scores by taking into account image difficulty as explained in the introduction. It is interesting to discuss the differences in the beta-weights between guns and knives in the image processing model. For guns view difficulty is weighted almost double compared to superposition. For knives, where superposition is weighted almost double compared to view difficulty, the contrary pattern was observed. We are currently conducting additional analyses to find out whether this effect is related to differential changes by 3D rotation. The reason why superposition is weighted much stronger in knives than in guns is probably due to the superposition formula which also reflects the size of the threat items. In the X-Ray ORT knives differ more in size than guns. Thus, the regression coefficient patterns reflect actual characteristics of the weapon categories. The scatter plots (Figure 6) reveal that, especially in knives, there is a certain ceiling effect. Therefore, it might be of value to use non-linear regression for modeling hit rates in the future. Apart from that, this study can be viewed as the basis for further statistical models for the prediction of individual screener responses to single x-ray images using binary logistic regression. In addition, together with the development of enhanced and additional image based predictors we intend to develop parallel statistical models to predict hit rates as well as false alarm rates.

6 Appendix

Clutter formula high-pass filter where f_x and f_y are its frequency components, f is its cut-off frequency and where d is its fall-off.

$$hp(f_x, f_y) = 1 - \frac{1}{1 + \left(\frac{\sqrt{f_x^2 + f_y^2}}{f}\right)^d} \tag{5}$$

This high-pass filter represents a 2-D matrix in the Fourier frequency space. Therefore an inverse Fourier transformation is applied to transform it into a convolution kernel in the spatial domain.

Acknowledgements

This research was funded by the European Commission Leonardo da Vinci Program (VIA Project). Thanks to Franziska Hofer and Diana Hardmeier (Department of Psychology, University of Zurich) for their contribution. Thanks to Christian Wallraven (Max Planck Institute for Biological Cybernetics, Tübingen) for valuable discussions regarding automated estimation of image-based factors.

References

GALE, A., MUGGLESTONE, M., PURDY, K., AND MCCLUMPHA. 2000. Is airport baggage inspection just another medical image? In *Medical Imaging: Image Perception and Performance. Progress in Biomedical Optics and Imaging*, vol. 1(26), 184–192.

GHYLIN, K. M., DRURY, C. G., AND SCHWANINGER, A. 2006. Two-component model of security inspection: application and findings. In *16th World Congress of Ergonomics, IEA 2006, Maastricht, The Netherlands, July*, 10–14.

GREEN, D. M., AND SWETS, J. A. 1966. In *Signal detection theory and psychophysics*, New York: Wiley, 187–194.

HARDMEIER, D., HOFER, F., AND SCHWANINGER, A. 2005. The object recognition test ort - a reliable tool for measuring visual abilities needed in x-ray screening. In *IEEE ICCST Proceedings*, vol. 39, 189–192.

KRUPINSKI, E. A., BERGER, W. G., DALLAS, W. J., AND ROEHRIG, H. 2003. Searching for nodules: What features attract attention and influence detection? In *Academic Radiology*, vol. 10(8), 861–868.

LIU, X., GALE, A., PURDY, K., AND SONG, T. 2006. Is that a gun? the influence and features of bags and threat items on detection performace. In *Contemporary Ergonomics, Proceedings of the Ergonomic Society*, 17–22.

MAHFOUZ, M. R., HOFF, W. A., KOMISTEK, R. D., AND DENNIS, D. A. 2005. Effect of segmentation errors on 3d-to-2d registration of implant models in x-ray images. In *Journal of Biomechanics*, vol. 38(2), 229–239.

MCCARLEY, J. S., KRAMER, A. F., WICKENS, C. D., VIDONI, E. D., AND BOOT, W. R. 2004. Visual skills in airport-security screening. In *Psychological Science*, vol. 15, 302–306.

SCHWANINGER, A., HARDMEIER, D., AND HOFER, F. 2004. Measuring visual abilities and visual knowledge of aviation security screeners. In *IEEE ICCST Proceedings*, vol. 38, 258–264.

SCHWANINGER, A., HARDMEIER, D., AND HOFER, F. 2005. Aviation security screeners visual abilities & visual knowledge measurement. In *IEEE Aerospace and Electronic Systems*, vol. 20(6), 29–35.

SCHWANINGER, A., MICHEL, S., AND BOLFING, A. 2005. Towards a model for estimating image difficulty in x-ray screening. In *IEEE ICCST Proceedings*, vol. 39, 185–188.

SCHWANINGER, A. 2004b. Computer based training: a powerful tool to the enhancement of human factors. In *Aviation Security International, FEB/2004*, 31–36.

SCHWANINGER, A. 2005b. Increasing efficiency in airport security screening. In *WIT Transactions on the Built Environment*, vol. 82, 407–416.

SCHWANINGER, A. 2006b. Airport security human factors: From the weakest to the strongest link in airport security screening. In *Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 - December 1*, 265–270.

SLUSER, M., AND PARANJAPE, R. 1999. Model-based probabilistic relaxation segmentation applied to threat detection in airport x-ray imagery. In *Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering*.

YING, Z., NAIDU, R., AND CRAWFORD, C. R. 2006. Dual energy computed tomography for explosive detection. In *Journal of X-Ray Science and Technology*, vol. 14(4), 235–256.