Searching for Possible Threat Items to Safe Air Travel: Human Error and Training

Xi Liu and Alastair Gale

Applied Vision Research Centre, Loughborough University, Loughborough,
LE11 3UZ, UK
{x.liu3, a.g.gale}@lboro.ac.uk

Abstract. An eye-tracking experiment is reported which investigates the underlying factors that affect training in the visual search of air passenger luggage for possible threat items so as to reduce errors and improve safe air travel. In this study, naïve observers learned to search for terrorist threat items of guns, knives and improvised explosive devices (IEDs) in airport passengers' X-ray luggage images. During training, each participant viewed the same number of learning trials of guns, knives or IEDs. Transfer performance was measured in a same search task in which each participant was more familiar with the visual appearance of half of the test targets. Detection performance and eye movement data both showed improvement in the efficiency of search and recognition with practice, while the skills were stimulus-specific so that performance was degraded when novel targets were introduced. Perceptual learning and human errors of the implications for screener training are discussed.

Keywords: visual search, perceptual learning, airport X-ray luggage image.

1 Introduction

Airport security screeners regularly visually search X-ray images of passengers' luggage for potential threat items. A large number of funds have been invested for enhancing aviation security since the events of 9/11, including new screening equipment, technologies and security personnel training. To summarize the new security screening techniques are: dual-energy analysis for estimating the atomic numbers of objects in passenger bags; transmission scattering and computer tomography to separate objects in complex backgrounds; and scattered X-ray energy imaging techniques for detecting plastic explosives better [1]. Substantial enhancements are required to the existing techniques to meet reliable detection performance and acceptable speed. The report about aviation security (for the U.S. Government Accountability Office) shows that explosive detection systems had not been utilized for 100 percent of baggage because of the shortage of security screeners or some other reasons so that this task still relies heavily on the human interpretation of X-ray luggage images to detect terrorist threat items. However, this demanding task is accomplished in a few seconds and the images are difficult, having a low signal-noise ratio coupled with potential targets having unknown shapes in a cluttered

D. Harris (Ed.): Engin. Psychol. and Cog. Ergonomics, HCII 2007, LNAI 4562, pp. 750–759, 2007.

background. Decisions of false-negative (FN) and false-positive (FP) are inevitable in examining X-ray luggage images for terrorist threat items due to the innately human limitations of visual cognition. Reasons for mistakes could be failures of expertise or a lack of expertise [2]. Efforts to update and maintain the expertise of security screening, overcoming the limitations of human perception and cognition, are therefore a priority for safer air travel.

An obvious difference between an expert and novice is that the expert has more experience of objects in their domain. In a related research domain to airport security screening, research has supported that the minimum number of cases for a radiologist to interpret in a fixed period is necessary [3]. It was indicated that one reason for the performance of residents to be worse than mammographers was the lack of visual recognition skills which could be obtained from perceptual learning with computer-assisted feedback. Massed practice with feedback is very helpful; not only for improving sensitivity but also for some non-specific aspects, such as conceptual knowledge or a search pattern which can transfer to other unfamiliar tasks [4]. For many tasks, human visual performance improves dramatically with practice while learning does not transfer when conditions changes. Karni and Sagi [5] have found that learning in texture discrimination tasks is local and transfer does not occur when background elements rotate 90°. In the task of discriminating the offset sign in a vernier discrimination task [6], learning does not transfer from horizontal orientation tasks to vertical orientation tasks or vice versa. For some situations, results have shown that a harder training phase induces learning which transfers to novel stimuli [7]; exposure of easy stimuli facilitates learning in difficult trials [8] and the ability of ignoring useless information from task-redundant information transfers to new stimuli [9]. In the task of searching for knives in passengers' X-ray luggage images, part transfer was observed from eye movement data as participants could fixate on targets more quickly after training [10].

The experiment reported here assessed the effect of practice with immediate object feedback and measured the transfer of perceptual learning in order to reveal how knowledge of security screening is organized. Participants practiced search and recognized threat items (guns, knives and improvised explosive devices (IEDs)) for three sessions in one day, which were displayed in a balanced order for each participant. Participants were tested on new stimuli with the same task following practice, and on the next day, so that transfer was measured by comparing performance on the trained and new stimuli. It was assumed that participants were more familiar with half of the test stimuli than the other half since half the test stimuli were scored with a higher similarity with practice stimuli than the other half in a pilot study, which gave a chance to evaluate the degree of transfer. Eye movements were recorded in this experiment for further analysing error reasons and revealing the mechanisms of performance improvement with practice.

2 Method

2.1 Participants

Twelve naïve people (6 female) took part in this experiment. Participants' visual acuity was normal or corrected-to-normal.

2.2 Stimuli

Experimental targets were eight guns from two viewpoints, eight knives and eight IEDs, which were all chosen from a large set of threat items. These items were scored for visual similarity (defined here as "similar objects share visual characteristics such as colour, texture, size, orientation and shape") between objects of the same kind (e.g. a gun was only compared with a gun, not with a knife or an IED) in a pilot study. Four sets of two guns from two viewpoints, two knives and two IEDs each, termed Set 1, 2, 3 and 4, were derived so that the visual similarity within a set was scored higher than the visual similarity between sets. Moreover, there was high visual similarity between Set 1 and Set 3, and between Set 2 and Set 4; while low visual similarity existed between Set 1 and Set 2, and between Set 3 and Set 4.

Target-present images were generated by inserting target images into normal bag images randomly. There was only one threat item in each luggage image. The target objects of Set 1 and Set 2 were inserted into 60 normal bags for training purpose. Set 3 and Set 4 were inserted into 16 normal bags for test purposes. Target-absent images were 120 normal bag images without any threat items. Each session was composed of 16 target-present luggage images and 24 target-absent images.

2.3 Design and Procedure

The participants' task was to search for threat items (gun, knife or IED) in passengers' luggage images. Half of the participants were assigned to the group that learned targets from Set 1, and half were assigned to the group that learned targets from Set 2. On the first day, each participant completed three practice sessions (session 1, 2 and 3) and one test session (session 4). On the second day, the test session was done again (session 5). On each trail of training sessions, images were presented against a white background on a computer monitor for an unlimited time. Participants pressed a spacebar to indicate that they had finished searching and made their decisions using a five-point rating scale. If their decisions were higher than 2 ('probably absent'), they also had to indicate the location of a potential threat item. Then an image in which the target was clearly displayed for target-present stimulus was followed for unlimited viewing. In the test sessions, other than no feedback provision the procedure of search and decision-making was the same as in the training sessions. There were three minutes break between training sessions and ten minutes break before testing on the first day.

In order to eliminate the effect of session order on measurement of detection performance, the order of practice sessions was counterbalanced so each session appeared equally at each stage and equally before and after every other session. Since half the test targets were from Set 3 and half were from Set 4, participants might be more familiar with half of the test targets than the other half but all test targets were novel to participants. The composition of an IED was explained to participants before the experimental sessions. Before the test session, participants were told the shapes of target objects in the test sessions were different from training targets and feedback images were not available.

X-ray luggage images were displayed on a 21-inch (53 cm) 1280 × 1024 monitor and viewed from a 70 cm distance. Eye position was calibrated before each session and eye movements were recorded using a Tobii X50 eye tracker.

3 Results

In this study, dependent variables were analysed by two-way mixed analyses of variance (ANOVA) with sessions as a within-subject factor and sets as a between-subject factor. All participants' data were pooled together as the differences among the sets were not investigated in this study. Location of a threat item was considered so that false location with a positive response was considered to be a false negative decision, only correct location and positive response was scored as a true positive decision.

3.1 Performance Analysis

Decision confidence data was analysed using the Receiver Operating Characteristic (ROC) method using the software ROCKIT [11]. The mean overall performance of each session was expressed as an A_z value, the areas under the ROC curve, which jointly considered hits and misses. Figure 1 is the graph with A_z of session 1 as the X-coordinate and A_z of session 3, 4 and 5 as the Y-coordinate. This shows the performance variation of participants in the last training session and transfer sessions. The line is a reference line to the first training session. Points in the upper area of the line showed better performance than session 1, while points in the lower area of the line showed worse performance than session 1. Figure 1 intuitively shows that the performance of session 3, the last training session, was much better than session 1 where all of the points are in the area of the upper line - performance increased with practice. About half of the points in transfer sessions were in the lower area of line which indicated transfer performance was even worse than session 1 when novel targets were introduced.

An analysis of variance (ANOVA) revealed that the improvement was significant with practice, $F (2, 20) = 23.829$, $p < .001$, reflecting an increase of the overall hit rate from .71 of the session 1 to .89 of the session 3 and a decrease of the false alarm rate from .33 of session 1 to .08 of session 3. Analysis showed that there was no difference of performance between session 1 and transfer sessions, while performance of session 3 was significantly better than transfer sessions, $F (2, 20) =50.150$, $p < .001$ of session 4 and $F (2, 20) = 28.779$, $p < .001$ of session 5.

Figure 1 shows that the detection performance of some participants in the transfer sessions was worse than their performance in session 1. In order to investigate this interesting phenomenon, the hit rate in session 1 was divided into two parts since immediate feedback was provided and each threat item was displayed twice in different viewpoints and backgrounds in each practice session so that the detection performance would be enhanced due to the first presentation of targets. The hit rate of threat items in the first presentation (H_{first}) in session 1, similar targets ($H_{similar}$) and unfamiliar targets ($H_{unfamiliar}$) in test sessions were calculated separately (see Table 1). The $H_{similar}$ of session 4 and 5 were both better than H_{first} [not significant]. Hit rates decreased significantly as novel targets were introduced, $H_{unfamiliar}$ of session 4 and 5 were worse than H_{first}, F (1, 10) =68.820, p < .001, and F (1, 10) =30.179, p < .001 respectively. However, false alarm rates significantly decreased from .33 of session 1 to .13 of session 4, F (1, 10) =19.929, p = .001; and .14 of session 5, F (1, 10) =20.180, p = .001.

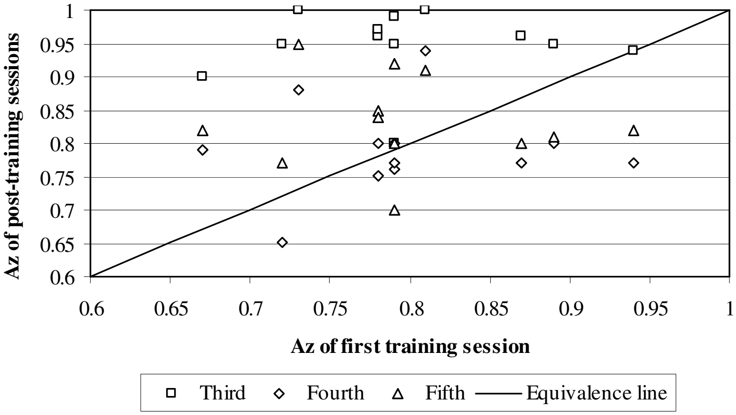

Fig. 1. Graph shows A_z values of participants in session 3, 4 and 5, as compared with session 1. The line is an equivalence line. Points in the upper area of the line represent performance better than session 1, while points in the lower area of the line represent performance worse than session 1.

3.2 Decision Time

Decision time decreased reliably with practice, F (2, 20) = 35.742, p < .001 of the target-absent images and F (2, 20) = 19.928, p < .001 of the target-present images. For transfer sessions, decision time of target-absent images were shorter than session 1 and longer than session 3, while decision time of target-present images was longer than all practice sessions. More details are presented elsewhere [12]. Decision time of false decisions was longer than that of true decisions for both target-present images and target-absent images in all training and transfer sessions (see Table 2), and differences in each session were significant, p < .05.

The decision time for the first presentation targets in session 1 was longer than that of similar targets in session 4 and session 5 (see Table 1), F (1, 10) =11.055, p = .001 and F (1, 10) =4.403, p < .05, respectively. There were no differences between decision times of first presentation targets in session 1 and decision times of unfamiliar targets in session 4 or session 5. Participants took more time to make decisions on unfamiliar targets than similar targets, F (1, 10) =18.107, p < .001 for session 4, and F (1, 10) =9.723, p = .002 for session 5.

3.3 Eye Movement Data Analysis

Eye movement data analysis further revealed mechanisms of performance change with practice and what happened in transfer. In consistence with the tendency of decision time decreasing with practice, the fixation numbers on images decreased reliably from session 1 to session 3, F (2, 20) = 14.782, p < .001 of target-present images and F (2, 20) = 8.169, p < .01 of target-absent images. The fixation numbers on target-absent images were significantly more than that on target-present images, p < .05. Participants took more time and fixated on more places in target-absent images than target-present images.

Table 1. Mean performance and eye movement data for targets in session 1, 4 and 5

			AOI		
Target category	Hit rate	Decision time (ms)	Fixation number on AOI	Time to first enter AOI (ms)	Dwell time on AOI (ms)
Session 1					
First presentation targets	0.65	14410	10	2307	3343
Session 4					
Similar targets	0.68	10843	10	1183	3874
Unfamiliar targets	0.28	16326	13	1033	4806
Session 5					
Similar targets	0.70	11149	10	1127	3577
Unfamiliar targets	0.38	16494	12	943	4601

The area of interest (AOI) was defined to analyze how participants visually processed the target area [13]. With practice, participants were inclined to focus on a threat item area (AOI) quickly (see Table 2), F (2, 20) = 4.432, p < .05. In the mean time, participants took less time on the AOI, F (2, 20) = 4.045, p < .05. Also the fixation numbers in the AOIs decreased, F (2, 20) = 3.298, p = .058. Participants could fixate and recognize targets quickly after training. In all practice sessions, the fixation numbers on AOIs of FN responses were less than that of TP responses, p < .05; the dwell duration on AOIs of FN responses was shorter than that of TP responses, p < .05; the time to first enter an AOI of FN responses was longer than that of TP responses, but not significant (see Table 2). To summarise, participants fixated on potential target areas of FN decisions with less fixation points and shorter dwell duration than that of TP decisions in practice sessions. Although participants needed more time when they made a FN decision than a TP decision, they fixated on the AOI of TP decisions longer than on AOIs of FN decisions. Some features of targets of TP decisions attracted participants' attention.

In the transfer sessions, fixation numbers on target-present images were more than that on target-absent images, F (1, 10) = 17.358, p = .002 for session 4 and F (1, 10) = 4.687, p < .05 for session 5. The dwell time on AOIs of similar targets and unfamiliar targets in session 4 and 5 all were longer than that of the first presentation targets in session 1 (see Table 1), but not significantly so. The time to first enter an AOI in session 4 was shorter than that of the first presentation targets in session 1, F (1, 10) =5.015, p < .05 for similar targets and F (1, 10) =5.677, p < .05 for unfamiliar targets. Also the time to first enter an AOI in session 5 was shorter than that of the first presentation targets in session 1, F (1, 10) =5.305, p < .05 for similar targets and F (1, 10) =5.584, p < .05 for unfamiliar targets. This indicated that the sensitivity to threat items was improved after training; no matter whether the decisions were correct or not.

Table 2. Decision time (millisecond), and eye movement data according to decision responses during training and transfer sessions

Decision time and eye movement data according to decision response	Training and transfer sessions				
	Session 1	Session 2	Session 3	Session 4	Session 5
Decision time					
False-negative (FN)	18914	8155	7847	14553	17352
True-positive (TP)	10945	6646	5323	12565	11427
False-positive (FP)	18459	16286	12495	16468	22436
True-negative (TN)	13491	8924	7356	9361	9436
Time to first enter AOI					
False-negative (FN)	2252	1322	1700	1112	1248
True-positive (TP)	1131	894	670	1011	854
Dwell time on AOI					
False-negative (FN)	2761	1848	1191	4657	3967
True-positive (TP)	3481	2779	2254	4894	4517
Fixation numbers on AOI					
False-negative (FN)	8	5	3	11	11
True-positive (TP)	10	8	7	13	12

3.4 Error Reasons and Skills Retention

The false-negative errors were classified into three categories: search error, recognition error and interpretation error [14]. If a target area was not fixated by any fixation points, then such miss responses were scored as a search error. If fixations or cumulative clusters hit on threat areas, and the gaze duration was less than 1000 ms, then these miss responses were termed a recognition error. If fixations or cumulative clusters hit on threat areas and the gaze duration was longer than 1000 ms, then miss responses were scored as an interpretation error. Other than session 3, the threat items were missed mainly due to interpretation errors (see Table 3).

Table 3. Missed errors in training and transfer sessions

Experiment sessions	Total number of missed error	Percentage of three types of missed errors (%)		
		Search error	Recognition error	Interpretation error
Session 1	54	15	11	74
Session 2	24	8	21	71
Session 3	22	9	50	41
Session 4	96	7	13	80
Session 5	88	8	11	81

The performance of session 5 was significantly better than that of session 4, F (1, 10) =9.181, p < .05, reflecting an increase in overall hit rates from .48 to .54. There was no difference in false alarm rate and decision time between session 4 and session 5. Fatigue effects should be considered in session 4 which was completed after three practice sessions. However, these data still show that participants retained knowledge and skills very well; even better after one day.

4 Discussion

In this study, perceptual learning, transfer and error reasons were investigated in a simulated airport security screening task. Not surprisingly, hit rate, false alarm rate and reaction time improved with practice. Moreover, eye movement data revealed that participants could fixate on targets and process them more quickly after practice. Observers got some perceptual experience with the appearance of targets from immediate object feedback so that they could detect and recognize target objects more quickly and accurately in the following training sessions. However, the improvement was stimulus-specific which was not maintained when novel stimuli were introduced. Hit rate on similar targets declined to the level of session 1, while hit rate on unfamiliar targets was even worse than on session 1. The benefits of stimulus-specific learning were only evidenced as the sensitivity on targets in transfer sessions was higher than session 1: participants were faster to fixate on familiar and unfamiliar target areas.

Eye movement data analysis showed that most of the targets in transfer sessions were missed due to interpretation errors, which indicated that observers fixated on the correct locations for a long duration but they still did not recognize targets. The poor performance of naïve people was caused by their lack of expertise in the airport security screening domain. In the task of searching for terrorist threats in passengers' luggage images at airport, expertise of screeners includes knowledge of threat objects, generic knowledge about X-ray images, the ability to deal with time pressure, a high vigilance against any suspects, and so on. In the simulated task, only knowledge about X-ray images and threat items was required, which indicated that recognition ability was a result of the experienced accumulation of reading X-ray luggage images. This rule could be applied not only to novices but also to experienced screeners according to the possible error types [2]. The performance of experts will decline to novices' level when their repertoire of rules is exhausted for solving novel situations in a familiar task. The main difference between experts and novices lies in the skill-based level and the rule-based level such that experts arrange attention and apply skills more effectively than novices.

In a study of searching for knives in X-ray luggage images, McCarley and colleagues [10] thought that familiarity with stimuli and the task led observers to fixate target area sooner which was considered as a proof that practice might not improve search skills. In our study, search errors and interpretation errors decreased with practice showing that performance improvement was the result of gaining search skills and object knowledge. When new stimuli were introduced in the test sessions, the number of search errors in the test sessions was less than session 1 even if the number of misses in the test sessions was much more than that of session 1.

Moreover, the time to first enter the AOI of the test sessions was significantly shorter than that of session 1. Effectiveness of search was improved with practice and partly transferred to new stimuli.

In the training sessions participants consistently took a longer time and made more fixation points on target-absent images than on target-present images. A serial search model could be used to interpret this result which also appeared in the previous study [15]. When participants implemented the visual search task, they kept searching at a certain rate until the target was found so that search might be terminated in the middle of inspection. Otherwise, they would continue searching until every object in the image was scrutinized. The longer decision time of target-present images in the test sessions demonstrated that stimulus-specific with object feedback training helped observers develop perceptual sensitivity of threat objects but that the decision time was affected by unfamiliar targets. Moreover, general knowledge about the features of X-ray images obtained from practice was very helpful to reject distractors so that target-absent images were examined quickly in the test sessions.

In summary, frequent exposure of stimuli with immediate feedback in a real visual search task is an effective training method to integrate general knowledge of X-ray luggage images and recognition ability into perceptual experience. Learning in the visual search of threat items is stimuli specific, such that screener training should enlarge knowledge of terrorist threats. Practice decreased search error rates which improved the effectiveness of search. Nodine and Kundel [16] modelled visual scanning patterns in radiology which was composed of a rapid global scanning process and a systematic focal recognition process. This showed that there were differences in the visual scanning patterns of searching mammograms between experienced and inexperienced readers [17]. These results indicated that visual scanning patterns would be changed with training.

Further research work of ours will examine the effect of computer-assisted visual feedback training in the domain of X-ray luggage image inspection. We argue that human errors should be viewed not only in considering the individual screener's performance but also through taking the system approach [18]. An inspection error may happen even though high technology systems have many defensive layers.

Acknowledgements. This research is supported by EPSRC and carried out in collaboration with QinetiQ.

References

1. Muthukkumarasamy, V., Blumenstein, M., Jo, J., Green, S.: Intelligent Illicit Object Detection System for Enhanced Aviation Security. In: International Conference on Simulated Evolution and Learning (SEAL'04), Busan, Korea (2004)
2. Reason, J.T.: Human Error. Cambridge University Press, New York (1990)
3. Nodine, C.F., Kundel, H.L., Mello-Thoms, C., Weinstein, S.P., Orel, S.G., Sullivan, D.C., Conant, E.F.: How experience and training influence mammography expertise. Academic Radiologys 6, 575–585 (1999)
4. Sowden, P.T., Davies, I.R.L., Roling, P.: Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity? Journal of Experimental Psychology: Human Perception and Performance 26, 379–390 (2000)

5. Karni, A., Sagi, D.: Where practice makes perfect in texture discrimination: Evidence for primary visual cortex plasticity. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 88, pp. 4966–4970 (1991)
6. Poggio, T., Fahle, M., Edelman, S.: Fast Perceptual Learning in Visual Hyperacuity. Science 256, 1018–1021 (1992)
7. Doane, S.M., Alderton, D.L., Sohn, Y.W., Pellegrino, J.W.: Acquisition and transfer of skilled performance: Are visual discrimination skills stimulus specific? Journal of Experimental Psychology: Human Perception and Performance 22, 1218–1248 (1996)
8. Ahissar, M., Hochstein, S.: Task difficulty and the specificity of perceptual learning. Nature 387, 401–406 (1997)
9. Haider, H., Frensch, P.A.: The role of information reduction in skill acquisition. Cognitive Psychology 30, 304–337 (1996)
10. McCarley, J.S., Kramer, A.F., Wickens, C.D., Vidoni, E.D., Boot, W.R.: Visual skills in airport-security screening. Psychological Science 15, 302–306 (2004)
11. Metz, C.E., Herman, B.A., Shen, J-H.: Maximum-likelihood estimation of ROC curve from continuously-distributed data. Statistics in Medicine 17, 1033–1053 (1998)
12. Liu, X., Gale, A.G.: Search for threat items in X-ray luggage images: How visual skills and recognition ability develop with practice. In: Bust, P.D. (ed.) Contemporary Ergonomics, Taylor and Francis, London (In press)
13. Liu, X., Gale, A.G., Purdy, K., Song, T.: Is that a gun? The influence of features of bags and threat items on detection performance. In: Bust, P.D. (ed.) Contemporary Ergonomics, pp. 17–22. Taylor and Francis, London (2006)
14. Kundel, H.L., Nodine, C.F., Carmody, D.: Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Investigative Radiology 13, 175–181 (1978)
15. Wolfe, J.M., Cave, K.R., Franzel, S.L.: Guided search: an alternative to the feature integration model for visual search. Journal of Experimental Psychology: Human Perception and Performance 15, 419–433 (1989)
16. Nodine, C.F., Kundel, H.L.: The cognitive side of visual search. In: O'Regan, J.K., Levy-Schoen, A. (eds.) Eye movements: from physiology to cognition, pp. 573–582. Elsevier, Amsterdam (1987)
17. Krupinski, E.A.: Visual scanning patterns of radiologists searching mammograms. Academic Radiology 3, 137–144 (1996)
18. Reason, J.: Education and debate. Human error: models and management. British Medical Journal 320, 768–770 (2000)