APPLICATION OF SIGNAL DETECTION THEORY TO THE RECOGNITION OF OBJECTS IN COLOUR-ENCODED X-RAY IMAGES

VASSIAS VASSILIADES

A thesis submitted in partial fulfilment of the requirements of
Nottingham Trent University for the degree of
Doctor of Philosophy

October 2010

Acknowledgements

I would like to express my sincere gratitude to my director of studies, Professor Paul Evans for his guidance and advice in the preceding years. I would also like to extend special thanks to my supervisors, Professor Lloyd Kaufman and Dr. Jer Wang Chan for their consistent help and support during this research project.

My parents Michael and Elli Vassiliades deserve special mention for their help and support during my degree course. A particular thanks to Dr. Simon Godber for his help with the experimental software and Dr. David Downes for his work on the experimental KDEX scanner. Further thanks to Ms Neranga Abeywicrama, Mr Omar Abusaeeda, Ms Rachel Bolton-King, and Dr. Xun Wang for their help as research assistants. Kind regards to Dr. Dhavalkumar Thakker for his insightful comments and Ms Evangelia Botis for her assistance with some of the illustrations. Additionally, numerous thanks to friends and colleagues for your ideas and suggestions over the years.

Finally, this research was funded by a grant from the US DHS Transportation Security Laboratory's Human Factors Program. The author would like to personally thank the program's director, Dr Josh Rubinstein, for his unwavering support throughout the project. The author also acknowledges Professor Dick Lacey, of the UK HOSDB, for his support of this program of research and in particular for the provision of the experimental x-ray system used to produce the stimuli, on a permanent loan basis.

Abstract

Aviation security personnel encounter problems when interpreting x-ray images of hand luggage. This research seeks to determine whether the performance of the human operator can be improved, in terms of both reliability and accuracy, through the employment of a novel multiple-view x-ray imaging technique. Thus, a series of experiments were undertaken with the aim of providing evidence for the feasibility of using KDEX displays to aid in the recognition of threatening objects in airport carry-on luggage; and furthermore demonstrate the real-world value of this technique.

This thesis describes experiments comparing how introducing depth information affects the performance of aviation security personnel attempting to detect various weapons in x-ray images of hand luggage.

Specifically, multiple 2-dimensional (2D) x-ray luggage scans were acquired and processed to create the perception of 3-dimensionality (3D) in kinetic displays. These results were compared with weapon detection in standard static 2D scans of the same luggage. Threatening objects hidden in this luggage were more readily detected in kinetic 3-dimensional images than in the standard images.

Initial results were obtained using greyscale images and limited to various types of knives. Subsequent experiments evolved to primarily use pseudo-colour x-ray images. These were encoded using the industry standard colours; orange for organic materials, green for mixture, and blue for metals. Additionally, as well as knives, the threat categories were expanded to include handguns, scissors and grenades.

In view of the remarkable improvement in performance afforded by the kinetic images future studies are warranted. For example, various features of the displays (e.g. differences among hues and contrasts) should be manipulated as parameters to determine if further improvements may be achieved. Additionally, different pseudo-colour schemes could be experimented with in order to potentially maximize the gains afforded by the kinetic display. Overall, when deployed in the field this technique may well revolutionize x-ray visual security screening.

Table of Contents

List of Figures

List of Tables

List of Acronyms

2AFC	Two-Alternative Forced Choice Design
2D	Two Dimensional
3D	Three Dimensional
fMRI	Functional Magnetic Resonance Imaging
HF	Human Factors
KDEX	Kinetic Depth Effect X-Ray
KM	Kinetic Monoscopic
MV	Multiview Monoscopic
OKR	Opto-Kinetic Response
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
RT	Response Time
SDT	Signal Detection Theory
SFM	Structure From Motion
SKE	Stereo Kinetic Effect
SM	Static Monoscopic
TIP	Threat Image Projection
UC	Uniform Connectedness
Y/N	Yes-No Design
zROC	z-Scores Receiver Operating Characteristic

1 Chapter One: Introduction

The experiments described in this thesis combine novel x-ray imaging techniques, previously developed by the university team, with well established psychophysical research techniques. Depth through motion, known as the kinetic depth effect (KDE), was incorporated with x-ray technology to produce displays that shall be referred to as “kinetic monoscopic” (KM). The purpose of this research was to compare these KM displays with the industry standard two-dimensional (2D), or “static monoscopic” (SM) displays (Vassiliades et al., 2008). Specifically, this thesis sets out to establish if, under laboratory conditions, KM displays offer a significant threat detection performance advantage, over and above currently standard techniques. Overall, this project forms part of a larger body of research that seeks to improve human reliability, and accuracy of detecting and classifying targets, through technological innovation, and thus promote the human operator model for security screening.

1.1 X-Ray Security Screening

X-ray based security screening was introduced in airports in the 1970’s following a series of high-profile hijackings where terrorists were able to smuggle weapons on board and take over the aircraft (Harris, 2002). The initial impetus for the development of these technologies was the 1972 deadline requiring the airlines to conduct total baggage screening for carry on luggage at USA airports. These early systems were relatively limited in their scope by today’s standards and properly deploying this equipment in the field followed a steep learning curve.

In the 1980’s, as equipment costs declined, x-ray scanners became more commonplace at airports across the world. Presently, in conjunction with more refined magnetometers (metal detectors), and explosives detection devices, x-ray scanners are considered the front line defence for security screening at airports and other sensitive installations, such as ports and public buildings.

Security scanners typically work by exposing luggage to low doses of x-ray radiation as they move on a conveyor belt through an inspection tunnel. Each bag is scanned line-by-line using a slit x-ray beam. The different objects inside the bags block varying amounts of x-ray beams depending on their material composition.

The pattern of x-ray beams, that traverse the luggage, is picked up by a linear detector array and stored digitally in an image buffer. The system then converts the accumulated x-ray signals into a line-scan 2D image (cf. Evans, 2002).

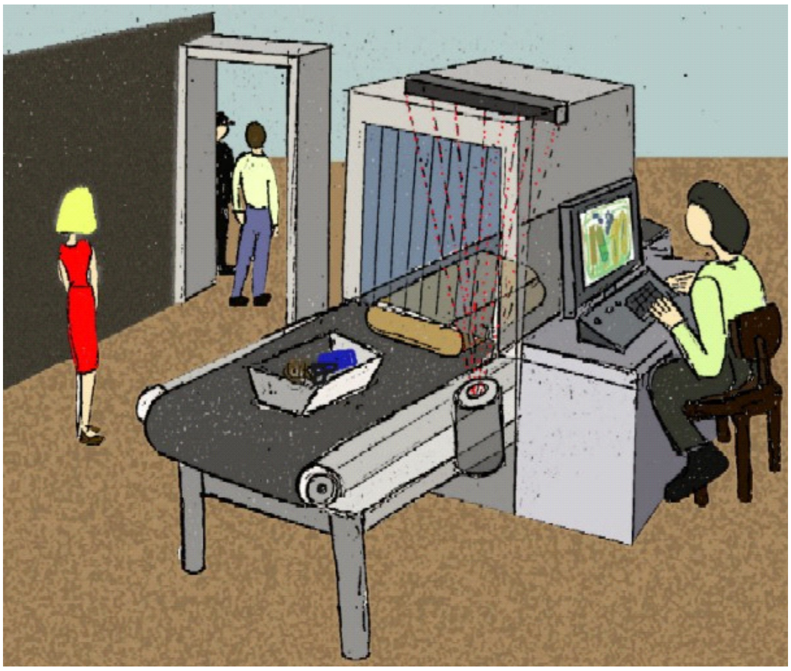


Figure 1. Typical airport security screening scenario.

The obtained image is typically displayed on a computer screen adjacent to the machine (see Figure 1). This enables the screener to non-invasively examine the entire set of luggage that passengers wish to take with them inside the airplane cabin for whatever items may be deemed prohibited by airport authorities, including possible weapons and explosives.

With the introduction of dual-energy x-ray systems, which permit a certain degree of materials discrimination (cf. Evans, 2002), the reliability of these systems increased. This is graphically presented to the operator as pseudo colouring of the images. The colouration of the various objects in the x-ray display depends upon the average atomic density of the underlying materials (see Figure 2). For instance, organic material such as paper, clothes and most plastic explosives appear orange. Metals, such as copper, are displayed in blue and opaque items (e.g. lead) are coloured in black. Intermediate density materials, such as dense plastics, are displayed in green. Finally, the intensity of the colour depends upon the average density or thickness of the material (cf. Sobania & Evans, 2005).

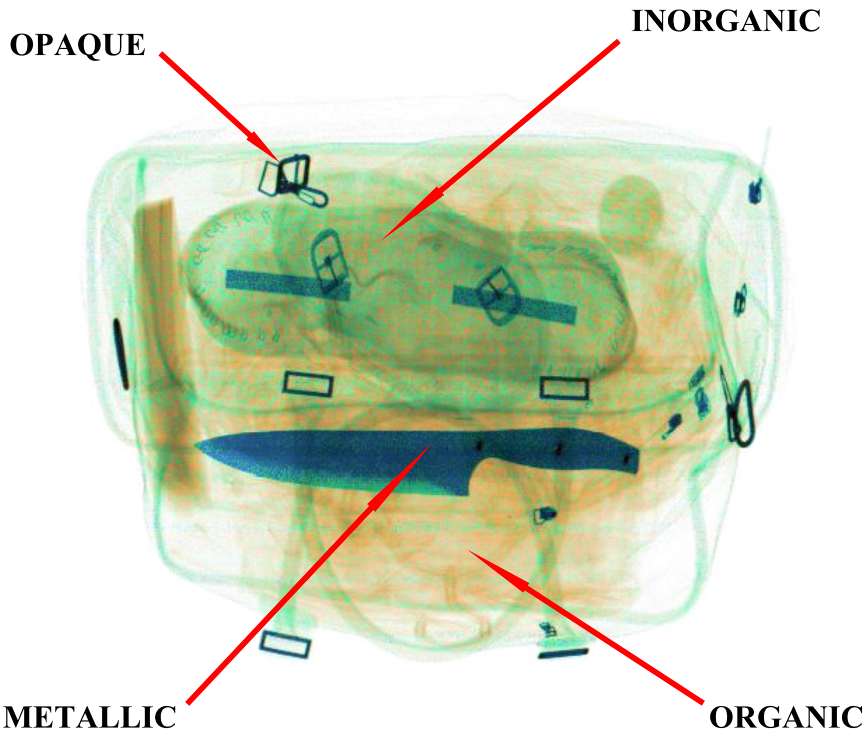


Figure 2. Example of a colour-encoded x-ray image.

However present day systems can still be defeated by a determined attacker. One notable example is the 1994 bombing of a passenger jet, en route to Japan from the Philippines, which resulted in the death of a passenger. The terrorist smuggled the components on board, hidden among normal travel items, assembled and planted the bomb under a seat (The Aviation Safety Network Website, Accessed 19 November 2008). The message here is clear. There needs to be an extensive evaluation of current technology and procedures, identification of the shortcomings and finally research aimed at finding possible solutions for each one of these.

1.2 Limitations of Current Technology

The most commonly used systems today consist of a linear x-ray sensor, which provides screeners with 2D line-scan images; colour-coded for average atomic number. The screening task includes several target categories (e.g. guns, knives, and explosives), and is compounded by the fact that the full member set of these categories is not known (McCarley et al., 2004). Lacking precise knowledge of the numbers and classes of potentially threatening objects, professional screeners must be able to decide if an object may or may not be a member of a class of objects that could conceivably pose a threat. Such a decision would warrant a more direct physical examination of the contents of the luggage. The screener must therefore rely upon his or her training and experience to visually sift through the cluttered x-ray images and make a decision as to whether the bags under inspection contain a threatening object or not.

However, this type of visual search strains the perceptual and cognitive capacities of the screeners (Harris, 2002). This process is exacerbated by the lack of visual cues to depth, an inherent limitation of the transmitted radiation technique used to produce these x-ray images. This complicates the screener's job, and frustrates human reliability in visual inspection tasks. Thus a new and better technique is needed that can re-introduce depth cues into the detection process.

1.3 Proposed Solution

The impetus for developing a new technique stems from the need to meet the ever increasing challenges of improving human-centred aviation security screening. In order to achieve this goal new technologies and procedures are needed (Fiore, Jentsch, & Rubinstein, 2004) that can improve human performance and reliability in x-ray security inspection tasks.

In the years since the 9-11 tragedy, computerized tomography (CT) scanners have been deployed with increasing frequency in the screening of checked luggage shipped in the aircraft's cargo bay (cf. Leone & Liu, 2005). While this expensive technology offers high fidelity images, it still remains problematic with respect to automatic threat detection. Due to their high purchase cost and large physical size of the device, CT scanners are as yet unfavourable to hand luggage screening. Additionally, CT scanners are comprised of mechanically complex rotational components. This means that, on top of the initial expense, the running costs of CT scanners also far exceed those of conventional 2D scanners. Thus a technology is needed that provides advanced 3D views, with closer to 2D costs; a niche that the proposed solution aims to fill.

The Kinetic Depth Effect X-ray (KDEX) technique works by scanning an object from successive angles, and thus obtaining a range of views from different perspectives (see Chapter 2). To achieve this, the system has a static configuration of multiple line-scan x-ray sensors, and uses a single x-ray source (Evans, 2002; Evans, 2003; Evans & Hon, 2002; Hon & Evans, 2002). Additionally, KDEX sequences may be produced in the same time period as conventional 2D x-ray images. Therefore this technique has the capability to replace existing 2D technology by providing additional 3D information at a fraction of the cost and bulk of CT technology (Vassiliades et al., 2008).

1.4 Research Questions

To summarise; in this thesis x-ray images, collected using the University group's new imaging technique, were subjected to a rigorous evaluation using experimental psychology techniques. Specifically, by employing the bias-free procedures of signal detection theory (Macmillan & Creelman, 2005) the hypothesis that the KDEX three-dimensional display provides a marked performance improvement over conventional static monoscopic two-dimensional displays was tested. The null hypothesis was that performance with standard static monoscopic (2D) may prove to be as capable as kinetic (3D) displays. To test these alternative hypotheses the following series of experiments were conducted (see Chapter 3), evaluating these two different image display modes.

Additionally, the effect, if any, of industry standard pseudo-colour encoding in detection performance for the kinetic display was briefly investigated.

1.5 Research Methodology

The research methodology for this project revolved around the following research activities: literature review, identification of the appropriate research design, experimental work and evaluation of the results with a subsequent refinement of the procedures.

1.5.1 Literature Review

In order to gain a better understanding of the broader research area, as well as keep up to date with current developments in the field, an extensive review of the literature was conducted. All the relevant topic areas were thoroughly examined both before, as well as concurrently, with the empirical phase of the project.

Security screening is at its core a visual search exercise; the screener is looking for a target. In order to accomplish that task he or she has to separate the various objects from the background; thus perceptually organise the scene. Then the relevant items of interest have to be located, which calls up the perceptual mechanisms of object identification. Additionally, since the screener is essentially looking for what is often a weak signal, signal detection theory becomes relevant as the framework of choice to address the research questions with. The new technique, evaluated in this thesis, is based on the kinetic depth effect, the underlying mechanisms of which have to be elucidated.

Finally, the experiments comprising this project were computer based. Consequently, as they involved human-computer interaction relevant considerations, for the experimental task itself, were examined during the preliminary work before the empirical section of the project.

1.5.2 Identification of the Appropriate Research Design

In consultation with the supervisory team the basic research design was formulated. This took into account the nature and availability of stimuli, as well as the type of research participants (inexperienced volunteers) available. Thus a Signal

Detection Theory (SDT) framework was selected as the most appropriate for this type of experiment. Two different SDT designs were evaluated in an initial pilot study and the Yes-No design prevailed. This then formed the basis for the design of subsequent experiments.

1.5.3 Experimental Work and Evaluation of the Results

Following the pilot study, a series of five further SDT experiments were undertaken in order to exhaustively test the experimental hypotheses. Considerable effort was put into the recruitment of volunteers, their training and the general execution of the research protocols. At the conclusion of each experiment extensive analyses of the collected data were conducted.

1.5.4 Refinement of the Procedures

The recorded data and behavioural observations made during each experiment, combined with the overall experience gained, were fed back into the design of subsequent experiments. Hence, where appropriate, changes were made to strengthen the design and improve the flow of the experimental protocols. Furthermore, with each successive experiment, as the knowledge base grew, so did the ability to correctly interpret the results.

1.6 Thesis Structure

Chapter 1 introduces the background to the research topic, discusses the current state of security x-ray screening, the limitations of this technology that spurred the impetus for the current project, and presents the research questions.

Chapter 2 contains literature reviews on the various topics of interest to this research. It includes discussions on signal detection theory (SDT), the kinetic depth effect (KDE), a description of kinetic depth effect x-ray images (KDEX), visual search, perceptual learning and object recognition. The chapter concludes with an overview of human factors (HF) considerations for this project.

Chapter 3 consists of a detailed overview of the design and results of the experiments conducted as part of this project. Experiments are presented in chronological order and each section contains a detailed description of the methodology and results, as well as a detailed discussion.

Chapter 4 summarises the experimental results, relates the experiments with each other, and draws conclusions on whether the project adequately addressed the research questions. Further discussion on the strengths and limitations of the project follows, as well as suggestions for further research.

2 Chapter Two: Literature Review

2.1 Kinetic Depth Effect

The literature review begins with a discussion of the kinetic depth effect. In order to better understand the underlying mechanisms the theoretical framework from which it originated first needs to be examined. This is followed by a discussion of the relationship between the kinetic depth effect (KDE) and stereopsis, and concluded with observations related to the KDE per se.

2.1.1 Depth from Motion

Visual parallax can provide useful information about distance. There are two types of parallax of concern here (Rock, p80, 1975). The first stems from the fact that the two eyes view the world from two different positions, which results in binocular disparity. This form of parallax leads to stereoscopic depth perception. The second type of parallax results from observer locomotion and head movement, which leads to so-called motion parallax. Helmholtz (1962) referred to this type of parallax as another important cue to depth.

Binocular disparity provides the human perceptual system with depth information because the aforementioned different positions of the two eyes results in a difference between the images formed on their retinas (Gogel, 1964). On the other hand, motion parallax is produced by the observer shifting his or her position thus changing the separation between the retinal images of objects located at different distances (Helmholtz, 1962; Rock, p80, 1975). The latter provides a powerful visual cue to depth (Helmholtz, 1962).

Wallach and O’Connell (1953) were the first to systematically investigate a special case of motion parallax, which they labelled the kinetic depth effect (KDE). In their pioneering research, they discovered that they could produce a very strong depth effect by casting shadows of fully or partially rotating objects on a screen (see Figure 3).

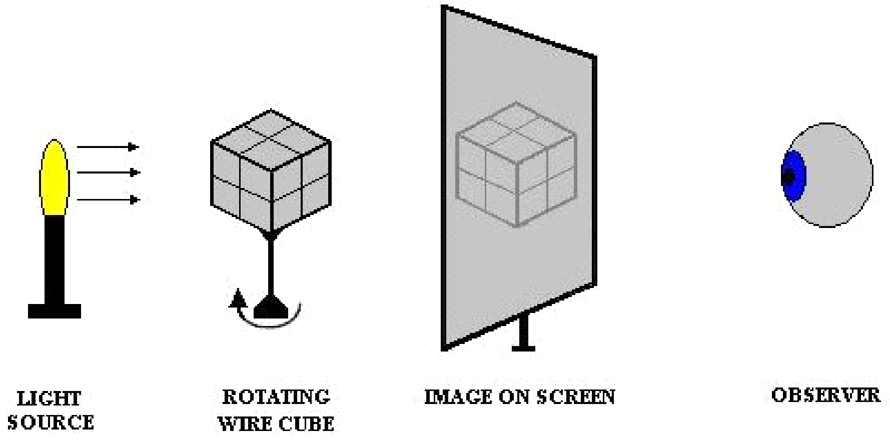


Figure 3. Producing the kinetic depth effect.

When properly oriented, with respect to the axis of rotation, the 2D shadows of contours of three-dimensional (3D) objects simultaneously change in both length and orientation (direction). An observer can then accurately deduce object shapes from the depth information in these shadowgraphs.

Previous work by Miles (1931), and subsequently by Metzger (1934), already demonstrated how shadows of objects in motion may elicit impressions of depth. However, these investigators did not identify the attributes of the two-dimensional shadowgraph display that are essential if one is to perceive the shadows as though they are rigid objects rotating in depth.

Returning to Wallach and O’Connell (1953), in their research they discovered that they could produce a very strong depth effect by casting shadows of fully or partially rotating objects so that their contours or edges simultaneously changed in both their length and direction on a display. An observer could then accurately deduce object shapes from the depth information in these shadowgraphs.

This phenomenon is also referred to as structure from motion (SFM) in as so far as the visual system extracts rigid three-dimensional structure from two-dimensional images that change in shape and size over time, thus producing depth cues (Ullman, 1979).

Gibson, (1950) described another form of motion parallax where the observer moves rapidly across a terrain, as in the popular example of a rail passenger looking out of the train's window. Nearer objects will appear to be moving faster than more far away objects; thus the speed of optic flow depends on each object's distance from the observer. The depth cue resulting from this difference in apparent speed is known as motion perspective. Kaufman (1974) suggested that since motion perspective essentially entails a semi-rotation of elements along an extended terrain, it can therefore be considered a form of KDE.

The mechanisms that underlie the perception of depth and structure from motion will now be examined. Most research into SFM and KDE has focused on full rotations of either shadows of wire frames projected onto a screen (e.g. Wallach & O'Connell, 1953; Gibson & Gibson, 1957) or luminous dot assemblies (e.g. Dosher, Landy, & Sperling, 1989; Caudek & Rubin, 2001; Fang & He, 2004). The latter, when set in motion, typically produce the three-dimensional shape of a cylinder.

From this type of research a lot can be learned about the perceptual mechanisms of the KDE. For example, Sperling, Landy, Dosher and Perkins (1989) reported that the kinetic depth effect is produced by automatic perceptual computations. These computations are based on velocity flow fields and the observers identified object shape in KDE displays by assigning different depth values to spatial locations of the moving 2D display in consistent proportion to their velocities. In subsequent research Sperling, Dosher, and Landy (1990) proposed that feedback is an essential element during the training of observers on how to interpret KDE displays. This ensures that observers learn how to make use of KDE-derived depth cues in experimental stimuli instead of simply generalizing from past experience to them. This last point is of interest in the future role of x-ray displays that incorporate the KDE.

Rock and Smith (1981) however caution that the perceptual system constructs the most reasonable representation of the presented stimuli. As such, they postulate that if the kinetic stimuli that lead to depth perception can be otherwise interpreted by the perceptual system, then these alternative transformations will be preferred and the perception of depth will not occur. This however is of minimal concern to the present research. Nevertheless, in general care should be taken during the preparation and display of motion stimuli so as to ensure that they are optimised to produce the KDE.

Semi-rotation (parallax) has also been clearly shown to create the perception of structure from motion (Durgin, Proffitt, Olson, & Reinke, 1995). Researchers have made progress over the years in the pursuit to elucidate the mechanisms of extracting depth cues from parallax. For instance, Nawrot (2003) reports that the perception of depth from motion parallax requires additional extra-retinal signals. These are provided by the slow eye movement system, termed the optokinetic response (OKR). This mechanism can supplement retinal signals and facilitate changing estimates of depth as a function of viewing distance.

The kinetic depth effect will now be compared with another type of depth producing motion. KDE is similar to the stereokinetic effect (SKE), whereby the rotation of certain types of figures creates a strong three-dimensional illusion (Wallach, Weisz, & Adams, 1956; Proffitt et al., 1992). Tauber and Kaufman (1977) added that stereokinetic effects are influenced by slow pursuit eye movements. On the other hand, it is important to state that the SKE differs from the KDE in some key aspects. Most important of all the SKE is in effect an illusion (Schmuckler & Proffitt, 1994) because no physical 3D object exists that can be deconstructed into 2D views based on the pattern used to create it. This is in contrast to the KDE which is created by projecting rotating 2D shadows of a 3D object on a screen, and not by spinning a 2D flat disk as in the SKE.

The ability to extract depth from motion seems to develop quite early. In fact, infants as young as five months old are known to be able to extract 3-dimensional information from KDE displays (Schmuckler & Proffitt, 1994), and like adults are sensitive to the between-contours, but not the within contours motion of the KDE (Proffitt et al., 1992).

The value of the KDE is further supported by animal studies where researchers demonstrated that monkeys were able to perceive three-dimensional structure from motion (Anderson & Siegel, 2005). Surely then the development of such a neurological mechanism for extracting depth from motion (cf. Paradis et al., 2000, for an fMRI study of the cortical basis of SFM) must serve an important evolutionary function to both human and non-human primates (cf. Tyler, 2004, for a discussion on the neural processing of moving stereoscopic forms in humans and monkeys). This phenomenon should be studied, and integrated into a broader technological and theoretical background, with the aim of improving the clarity, and hence the interpretation of x-ray displays.

2.1.2 KDE and Binocular Disparity

The focus of the discussion will now shift to the relationship between depth from motion and disparity. Overall it can confidently be stated that there is an essential isomorphism between KDE and binocular disparity (Rock, p114, 1975). For instance, research suggests that there is a psycho-anatomical link between the perception of depth from motion and the perception of depth from binocular disparity (Thompson & Nawrot, 1999). Furthermore, researchers have noted that viewing KDE displays produce the same type of binocular eye movements (i.e. vergence changes) when participants employed binocular disparity as a cue to depth (Ringach, Hawken, & Shapley, 1995).

In more recent work, Fang and He (2004) found that stabilized structure from motion, without disparity, induced disparity adaptation. More specifically their participants reported rotation after-effects during the experiments, which could only be adequately explained by direction-contingent disparity adaptation. Their findings suggest that the KDE stimuli, which have no disparity, activated disparity sensitive neurons. This they maintain provides evidence that stereo depth and kinetic depth may share a common early-stage neural mechanism in the visual system.

This type of adaptation was described much earlier by Wallach and Karsh (1963), who artificially induced a change in disparity by use of a telestereoscope. However, each eye's image of the same rotating object exhibited the same KDE, except that there was a distortion in depth introduced by the exaggerated disparity. Ultimately, the disparity cue was rescaled by the now conflicting KDE, so distortion was no longer perceptible. Even so, the amount of depth associated with the disparity was greatly altered.

One issue of concern is that KDE displays tend to be ambiguous since they are prone to perceptual reversals of the depth plane (Green, 1961; Day, 1965; Hochberg & Peterson, 1987). This depth reversal is similar to the classic optical illusion example of the Necker Cube (Necker, 1832). Motion perspective (Gibson, 1950) displays have also been shown to exhibit this phenomenon (Hershberger, Stewart, & Laughlin, 1976). Furthermore, Caudek and Domini (1998) found that the axis of rotation in SFM displays was systematically misperceived. This problem can be rectified by introducing additional perceptual information into the equation (Sperling & Dosher, 1994). For instance, Schwartz and Sperling (1993) report that adding simulated linear perspective alters the probability of reversals in the direction of rotation in KDE displays. Returning to Fang and He (2004), they were able to stabilize the perceived rotation of dots on a KDE rotating cylinder by introducing additional information (e.g. occlusion, disparity, speed of rotation etc) regarding the spatial context of the dots.

There is a sizeable amount of experimental work devoted to the synergy of depth from motion and binocular disparity (e.g. Rogers & Collett, 1989; Turner, Braunstein, & Andersen, 1997); however often with mixed results. For instance, Bradshaw, Parton and Eagle (1998) investigated the interaction of binocular disparity and motion parallax in depth and size perception. They report that when both cues are combined the accuracy of size and depth judgments increased, than when presented in isolation. This position is further supported by research findings which suggest that the process of three-dimensional structure recovery can influence two-dimensional motion correspondence in SFM displays (Mukai & Watanabe, 1999).

On the other hand, more recent research (Bradshaw, Parton, & Glennerster, 2000) suggested that the benefit of combining the two depth cues varies as a function of the type of visual task involved. It is possible then that when combining cues the stronger one (i.e. disparity) either overrides the weaker one (i.e. parallax), or that cues are processed separately and combined non-linearly depending on the task at hand (Hubona, Wheeler, Shirah, & Brandt, 1999). If the latter model is accurate then this selective fusion must occur at a higher-level of visual processing. This is because, as mentioned previously, there is compelling evidence that at a lower-level depth from motion and disparity may well share common neural mechanisms (Ringach, Hawken, & Shapley, 1995; Thompson & Nawrot, 1999; Fang & He, 2004). The implications are of course that either way caution should be exercised, when combining motion and stereo cues in computer generated displays, so as to achieve the desired outcome. This point will be readdressed later in the text.

In conclusion, after weighing all the evidence, it seems that for these purposes introducing disparity appears to be the best solution for minimizing spontaneous depth plane reversals in KDE displays.

2.1.3 Memory for the KDE

Finally, it is worth taking some time to discuss an interesting aspect of KDE reported over half a century ago. In a follow up paper, Wallach, O'Connell and Neisser (1953) describe a phenomenon they termed as the memory effect of the kinetic depth effect. In summary, they observed during their initial KDE experimental work that repeated exposure of their participants to structure from motion led them to subsequently infer that structure even when the objects were not moving.

Some indirect evidence comes from surface interpolation studies. For instance, Treue et al. (1995) found evidence that in KDE displays the observers' perceptual interpretation of a given object's boundaries influences the surface interpolation process, to a degree that up to a quarter of the stimulus could be absent without this hindering their interpretation of it. However other researchers (Jusczyk, Johnson, Spelke, & Kennedy, 1999) interpret this finding as evidence that observers may perceive object unity by detecting patterns of synchronous change, of which common motion can be considered a special case. So this memory for KDE may be a simple case of stimulus-specific learning.

An interesting possibility arises from work in stereoscopic apparent motion (e.g. Ito, 1999). For example, in their paper Fernandez and Farell (2006) describe conditions that produce a reversed structure from motion phenomenon. They were able to produce this motion-from-structure effect by having their participants simultaneously view two random-dot stereogram surfaces. However, interesting as this effect may be, it still fails to explain the phenomenon as it was described in the original context (cf. Wallach, O'Connell, & Neisser, 1953).

Overall direct evidence in the literature for this memory effect of the KDE is scant, and given the lack of any study replicating their findings, the author believes that what Wallach, O'Connell, and Neisser (1953) reported may have actually been an artifact of their procedures rather than an actual phenomenon. Some additional evidence comes from research that suggests that unless care is taken in the production of KDE displays then they can be contaminated by stimulus artifacts (Sperling, Landy, Dosher, & Perkins, 1989). Of course, the counterargument is that such artifacts are not as pervasive as originally proposed (Braunstein & Todd, 1990). Whatever the eventual conclusion may be it would be of particular interest if the present research uncovered any evidence to support the case for the memory for KDE as it was originally described.

2.2 Kinetic Depth Effect X-Ray Images

As is evident from the preceding discussion, motion can provide a powerful visual cue to depth, which greatly enhances the interpretation of spatially complex structures in shadowgraph images (Wallach & O'Connell, 1953). X-ray images are in essence shadowgraphs, because of the techniques used to obtain them, which makes this phenomenon of particular interest.

Taking advantage of these findings the University team has developed a technique which produces KDE image sequences from a single x-ray source and a static configuration of multiple x-ray detector arrays (Evans & Hon, 2002; Hon & Evans, 2002). The resulting "Kinetic Depth Effect X-ray" (or KDEX) imaging technique involves obtaining a range of views of the object under inspection (in this case various items of hand-luggage) from different perspectives in order to produce sequential parallax (Evans, 2003).

Furthermore, the multiple-view scanner produces images from what appears to be a sequence of discreetly positioned station points. This gives the observer the ability to essentially look around an object using these multiple views at different angles.

Additionally, by giving screeners the ability to view objects in their true 3D form it enables them to separate difficult to see targets from the background clutter.

2.2.1 Experimental Apparatus

The experimental multiple view dual-energy x-ray scanner (see Figure 4) employs a 140 kVP polychromatic x-ray source and a folded dual-energy linear x-ray detector array (see Figure 5). The x-ray beams emitted by the source are distributed in a conical shape. Bespoke collimation is used to convert the conical beam to a slit beam in order to reduce the amount of x-ray beams entering the inspection tunnel, as well as reduce x-ray scattering.

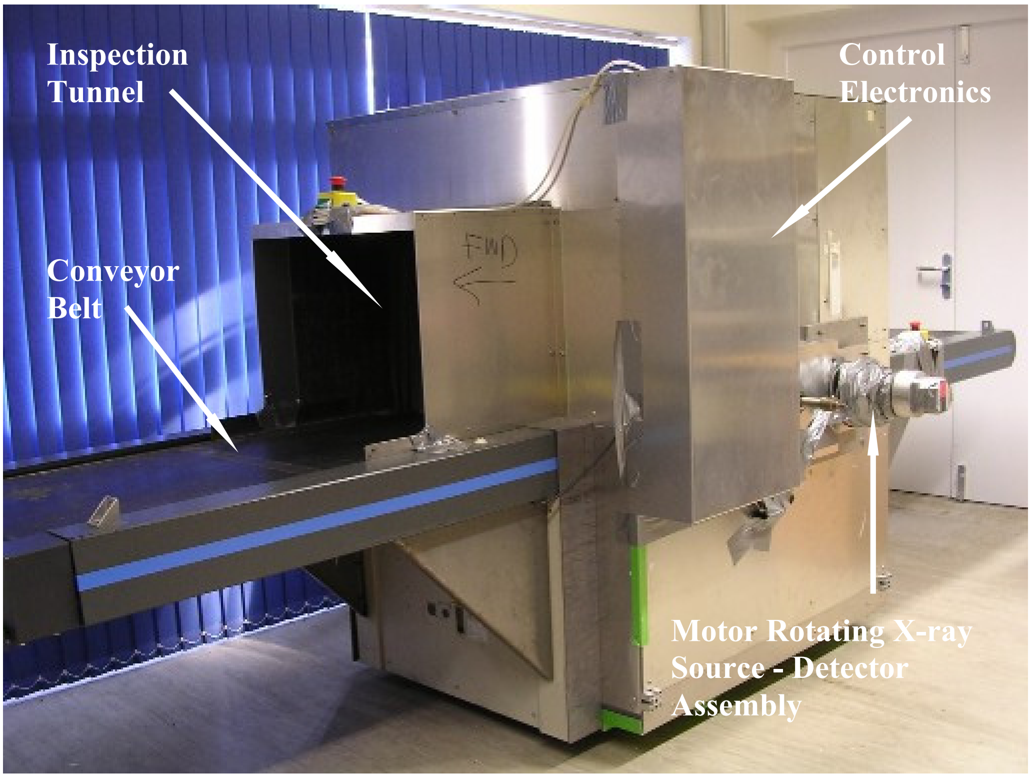


Figure 4. Experimental multiple view dual-energy x-ray scanner.

The dual-energy x-ray detector array consists of low-energy x-ray scintillators, and high-energy x-ray scintillators in a 'sandwiched' configuration, where a thin aluminium filter is placed between them. The front low-energy array is exposed to the full x-ray spectrum. Hence it absorbs the low-energy portion of the polychromatic spectrum. After passing through the filter, a significant amount of low-energy x-ray beams, as well as a small amount of high-energy x-ray beams, are removed from the spectrum. The remaining high-energy x-ray beams are subsequently detected by the rear high-energy array. This ‘sandwiched’ configuration allows both low-energy (~80keV) and high-energy (~140keV) x-ray data, of the irradiated object, to be acquired from a single exposure.

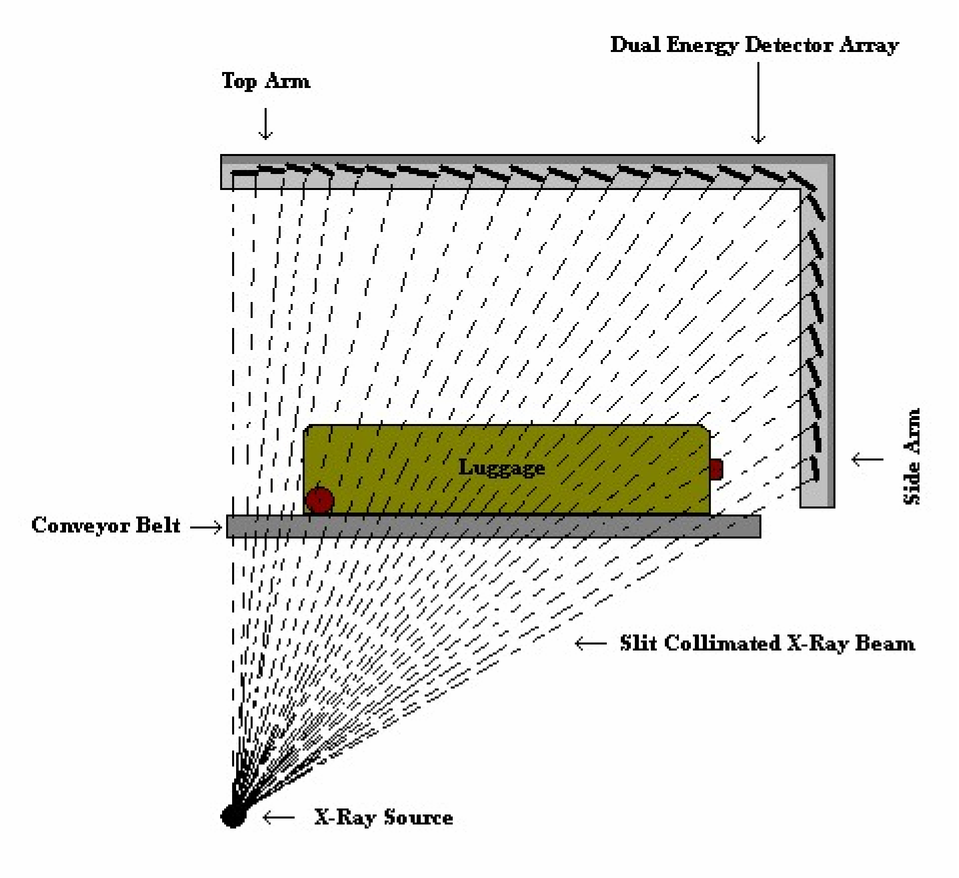


Figure 5. Folded dual-energy x-ray detector array.

Furthermore, the detector array is composed of two arms, in an L-shaped configuration, upon which the sensor modules are mounted (see Figure 5). The sensors in the top arm collect the data used to construct the top two thirds of each x-ray image, whereas the sensors on the side array collect the data for the bottom third.

2.2.2 Image Acquisition and Display

Each perspective view of the luggage is acquired using the slit x-ray beam in a manner similar to that involved in conventional 2D x-ray machines. As the bag is translated through the inspection tunnel, it is scanned line-by-line by the slit x-ray beam. The x-ray beams transmitted through the luggage are recorded by the linear detector and stored digitally in an image buffer. By cyclically storing the x-ray signals (i.e. luggage image information), a 2D line-scan image is produced.

After the scan is complete the conveyor belt returns the bag to the starting position. The entire x-ray source/detector assembly is then rotated, in relation to the conveyor belt, and the bag is re-scanned at the new angle. This enables the collection of multiple views at different angular positions. In this way the image output of a single pass multiple view scanner can be simulated.

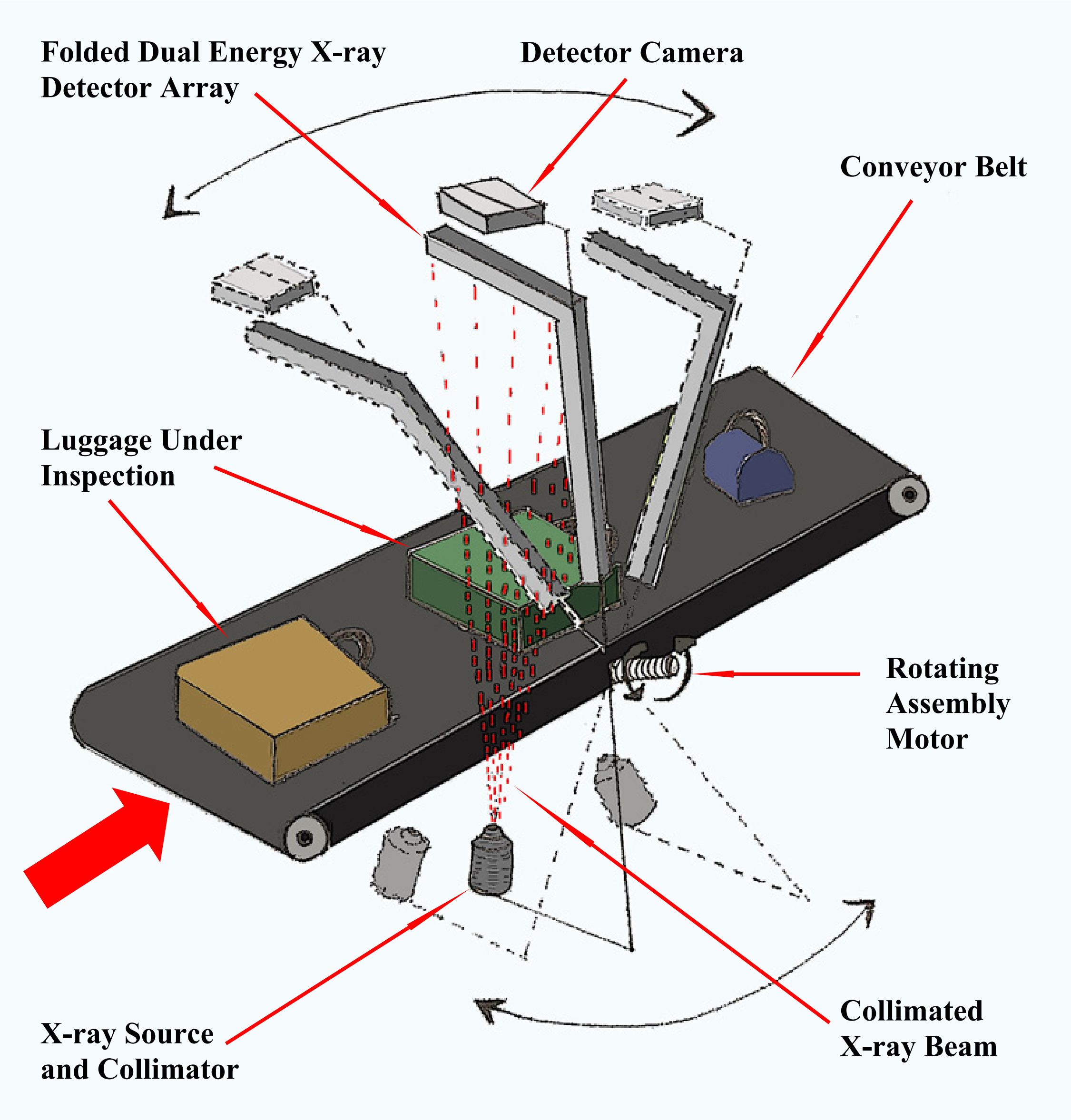


Figure 6. Kinetic depth effect x-ray system schematic.

The above schematic (Figure 6) illustrates the process of luggage inspection in the KDEX system (cf. Hon & Evans, 2002). The bag rides on the conveyor belt and transverses through the inspection tunnel. With each successive shift of the x-ray source/detector array angle, sequential multiple views are gradually obtained.

2.2.3 Colour Encoding

The low energy and a high energy x-ray data characterises the integrated effects of photoelectric absorption and Compton scattering. The difference between the dual-energy x-ray signals is exploited to broadly discriminate an inspected object into three material classes as a function of atomic number (Sobania & Evans, 2005). The organic class is for compounds with atomic number $Z \leq 10$, a mixture of organic and metallic material (or inorganic) class is for compounds with $10 < Z < 20$ and the metallic class is for compounds with $Z \geq 20$. It should however be noted that the labels "organic", "inorganic" and "metallic" are non-scientific. This terminology was adopted by the security industry, purely for the purpose of convenience, when describing objects in dual energy x-ray images.

The resulting discrimination information is presented to the human operators by colour-encoding the x-ray images using calibration lookup tables, and RGB colour palettes. Typically organic compounds are displayed in orange; metallic in blue and a mixture of organic and metallic in green (see Figure 2). This colour encoding scheme provides a general indicator of materials but is by no means a precise definition, as the degree of overlap of different materials can alter the coloration. Nonetheless, security x-ray screening systems utilising dual-energy materials discrimination are widely deployed in the major airports throughout the world.

2.3 Signal Detection Theory

This section initially describes the fundamentals of signal detection theory. This is followed by a discussion of receiver operating characteristic curves, and concludes with detailed examples of how this framework can be applied to this research.

2.3.1 Overview

Signal Detection Theory (SDT) was initially developed in the 1950’s (Harvey, 1992) and most of the initial work was in the area of radar detection. It provides researchers with a helpful framework for analysing decisions in terms of simple yes or no answers (Wickens, 2002).

The theory was first publicized by Tanner and Swets (1954), and its use in the field of experimental psychology was pioneered by Green and Swets (1966). In the intervening years it gradually gained acceptance and has become an invaluable tool in psychophysical experiments. Its particular strength lies in the ability to describe performance in ambiguous situations (Wickens, 2002), in the case for instance when dealing with a weak signal that might be confused with “noise”. Furthermore SDT has the ability to quantify observer bias and thus give a more accurate report on task performance; which is, in a sense, described as the observer’s sensitivity to the signal, independent of his or her bias.

In a typical signal detection trial a target may or may not be presented. For example, if a target is present on a specific trial, that trial is referred to as a “signal plus noise trial”, or simply as a “signal trial”. When a target is not presented during a particular trial, then that trial is referred to as a “noise” trial. This distinction is predicated on the now widely accepted notion that additive noise is an ever present condition. For example, in attempting to detect a weak acoustic stimulus, that stimulus must be distinguished from the pervasive background of environmental noise, as well as the ever present neural noise.

Given this distinction between signal plus noise trials, and noise trials it can be seen that the detection procedure is a discrete task, whereby in each trial a stimulus is presented and a response follows (Sperling & Dosher, 1986). Alternatively, the signal may not be present, but the observer is merely alerted that another trial is about to begin. In this case no signal at all is provided, but the participant must decide if a signal had or had not been presented during the interval of the trial. Hence, the participant must discriminate between signal and noise.

In many such experiments the signal plus noise and the noise alone trials are presented in random order but with equal probability. The typical procedure is to employ an equal number of signal and noise trials, thus maintaining a 1:1 signal to noise ratio (cf. Corwin, 1994, for a discussion of SDT models with unequal number of target and distractors). If the participant correctly detects the signal then this is considered a "hit" (H). However if he or she erroneously reports a signal when it is not present, then this positive response is referred to as a "false alarm" (FA). Conversely, if the observer fails to report a signal when it is present it is considered a "miss" (M). Finally, if the observer responds that no signal is present, when that is indeed the case, then the response is considered to be a "correct rejection" (CR).

Figure 7 (cf. Wickens, 2002; Macmillan & Creelman, 2005) graphically represents these response contingencies with respect to hypothetical signal and noise distributions. The top plot represents signal trials and the bottom plot noise trials. This example features the performance of a hypothetical participant in a SDT experiment, where the noise distribution has a mean of zero, and the signal distribution a mean of +1.

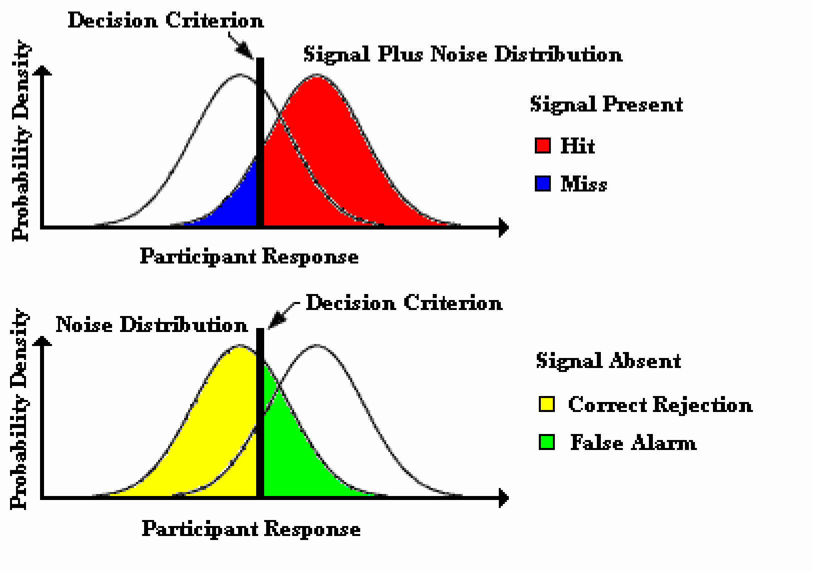


Figure 7. SDT Gaussian signal and noise distributions.

In this scenario the signal and noise distribution variances are equal to each other (or noise variance = signal variance = 1). However, in practice the noise and signal variances may be unequal, a condition that dictates the use of different statistics rather than the d' as measures of sensitivity (Wickens, 2002; Macmillan & Creelman, 2005).

Signal detection theory concerns itself with two aspects of an observer's response pattern, sensitivity and bias. Both these parameters are of interest when trying to paint a complete picture of performance in a task.

Sensitivity is expressed by determining the hit and false alarm rates and calculating the distance between the means of the two distributions. The derived measure d', referred to as the discriminability (or sensitivity) index, expresses the sensitivity of the observer, or how good he or she is at discriminating between noise and signal trials. For instance, in the hypothetical example depicted in Figure 7, d' = 1.

Generally, the more salient the signal is, then the higher the d' value should be. Presumably, when d' is zero then the threat is not at all salient, and observers cannot detect any difference between the two classes of stimuli (signal and noise). Their performance is then no different than chance. A d' value of 1 or higher indicates the opposite case where the observers exhibit increasingly better performance.

Response bias on the other hand is the degree to which each observer has a preference to respond affirmatively or negatively. This is especially relevant when dealing with human observers, who are after all individuals and as such despite receiving identical instructions are bound to differ amongst each other to a certain degree. It is therefore possible to have different observers who are similar in their sensitivity but differ in terms of bias (cf. Macmillan & Creelman, 2005, for more details on data pooling over different participants).

One commonly used bias index is the decision criterion (c) that the observer sets and divides the decision space (see Figure 7). The position of the so-called decision criterion may reflect a bias to respond either negatively or positively. This bias may be affected by the relative numbers of signal and noise trials in an experiment. However because of the way it is computed this depends solely on the false alarm rate (Wickens, p26, 2002). A better way of expressing observers' bias is to calculate the likelihood ratio (β); which expresses their tendency of saying yes or no and is a function of both the hit and false alarm distributions. However because β is inherently asymmetrical its natural logarithm (log β) is commonly used to express bias (Wickens, 2002).

The basic equations, used to calculate the SDT parameters described in the preceding text, are presented below (cf. Macmillan & Creelman, 2005; Wickens, 2002).

Table 1. *Basic SDT equations.*

Parameter Name	Equation
Hit Rate (HR)	HR = H / (H + M)
False Alarm Rate (FAR)	FAR = FA / (FA + CR)
Discriminability (or Sensitivity) Index	$d' = z_{(HR)} - z_{(FAR)}$
Decision Criterion	$c = -\frac{1}{2}[Z_{(HR)} + Z_{(FAR)}]$
Likelihood Ratio	$\beta = e^{d' \times c}$
Natural logarithm of the Likelihood Ratio	$\ln(\beta) = d' \times c = -\frac{1}{2}[Z_{(HR)}^2 + Z_{(FAR)}^2]$

Note. H = number of hits, M = number of misses, FA = number of false alarms, and CR = number of correct rejections.

A testament to the versatility of signal detection theory is its use as an analytical tool in many different contexts across the literature. For instance, Azzopardi and Cowey (1997) utilized SDT to probe the secondary visual channels in a person with blindsight, and Manning (1998) reports that SDT is the routine method for measuring diagnostic performance in medical radiography, where no two “signal” trials may be exactly alike. From studying anorexia nervosa (Smeets et al., 1999) to assessing race discrimination (Greenwald et al., 2003) SDT has proven its value as a research tool numerous times.

In conclusion, depending on experiment type, SDT can follow a detection, discrimination, or identification paradigm. A discussion will follow, later in the text, on specific scenarios for applying SDT procedures to the present requirements.

2.3.2 Receiver Operating Characteristic

Receiver operating characteristic curves, most commonly known as ROC curves are an important part of signal detection theory framework (Fawcett, 2006). The ROC curve was first developed during World War II for assessing the performance of radar operators in detecting enemy targets in the battle field (Green & Swets, 1966). Since then, ROC analysis has been widely used in medical radiology, experimental psychology and other areas for many years.

The ROC is essentially a plot of the hit rate (y axis) as a function of the false alarm rate (x axis). In general by observing the shape of the ROC an insight can be gained into the performance of a particular observer at a given task. More specifically, the closer the curve follows the left-hand and top borders of the ROC space, the more accurate the performance is at the task. On the other hand, the closer the curve comes to the major diagonal of the ROC space, the less accurate the performance. This is graphically illustrated in Figure 8.

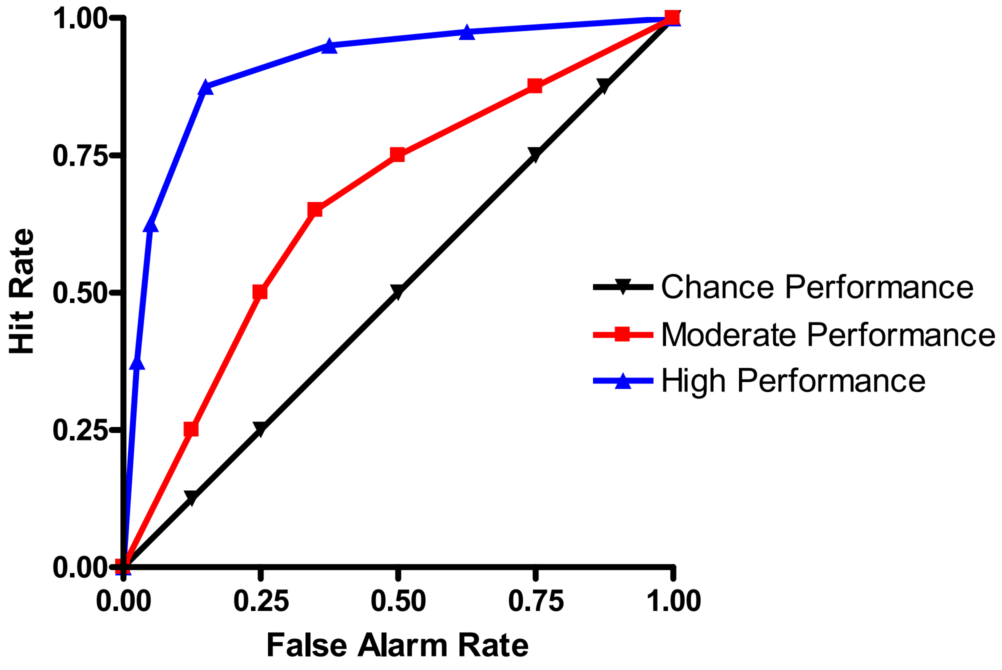


Figure 8. Examples of ROC curves.

The ROC can also be transformed into standard scores (or z-scores), in which case, if the underlying distributions are Gaussian, it will be a straight line. The corresponding equation then becomes $z_{(HR)} = z_{(FAR)} + d'$. The equal variances model predicts that the slope of this line should be 1, and the better the observer's performance, the higher the elevation of the line should be. A graphic example follows in Figure 9.

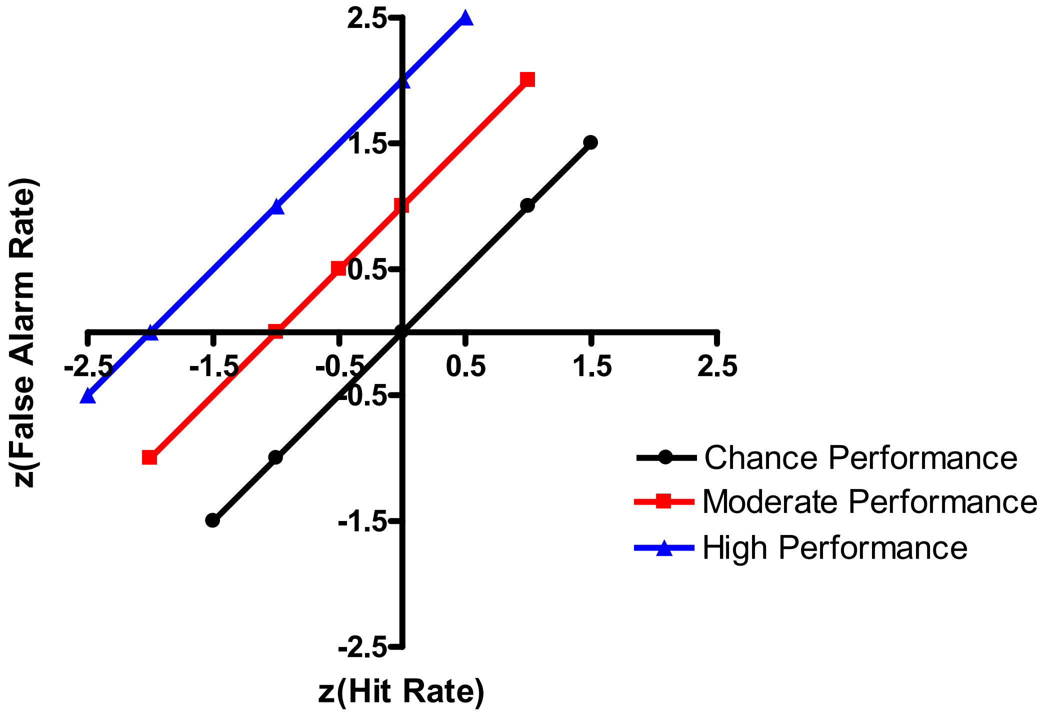


Figure 9. Examples of zROC lines.

However, in practice zROC slopes are often non-unit. In such cases several theorists have associated low sensitivity with slopes closer to 1, and also maintain that increased sensitivity results in increasingly shallower slopes (Green & Swets, 1966; Macmillan & Creelman, 2005). As will be seen later in the text, this is reflected in the data obtained during the empirical part of this project (see Chapter 3).

The area under the ROC curve (AUC or A_z), predicted by the Gaussian model using the equation $A_z = \Phi(d' /\sqrt{2})$ (where Φ is the normal distribution function), is also a practical performance measure (Hanley & McNeil, 1982; Marzban, 2004). It is especially useful in cases where unequal variances in the signal and noise distributions warrant against the use of other measures, such as the d' (Wickens, 2002; Macmillan & Creelman, 2005). However other theorists have however raised doubt about the underlying justification for using the area under the curve (AUC) measure. For instance Balakrishnan (2003) proposes as an alternative statistic to simply use the proportion of correct detection responses in a yes-no detection task of an unbiased decision-maker. Of course this latter point also implies that one should use 2AFC paradigm to limit the range of bias.

It is also worth noting that the area under the ROC curve (A_z) can never have a value greater than 1.0. Furthermore, when A_z is approximately 0.5, then sensitivity is essentially at the level of chance. So, if for instance $A_z = 0.98$, then observer performance is near perfect. By contrast, a similar level of performance represented by the d' must necessarily be > 3.0. This would indicate that the separation between the peaks of the hypothetical underlying normal distributions is at least 3 standard deviations apart; this is an enormous difference in discrimination.

Overall, there is general consensus in the literature on the usefulness of receiver operating characteristic methods. For instance, as aforementioned, it is the preferred method for measuring performance in diagnostic radiography (Manning, 1998; Obuchowski, 2003; Park, Goo, & Jo, 2004). Meteorologists are also interested in this procedure (Marzban, 2004), and SDT has received wide acceptance as an analysis tool in weather forecasting (Harvey et al., 1992). Another facility offered by SDT is the application of ideal-observer analysis to visual processing (cf. Geisler, 1989). This is of great importance because as will be seen later on it enables researchers to apply SDT models to visual search, which is a great benefit for the current research.

However, before concluding the discussion on signal detection theory it would be beneficial to briefly mention other relevant ROC methods. These techniques are derived from the classic ROC approach, used to describe individual observer performance on a detection task, and are typically employed in diagnostic radiology.

The “Multiple Reader Multiple Condition ROC” (MRMC) technique is a widely used assessment tool and involves designing studies where many different observers examine all available stimuli, within a certain dataset. For example a group of diagnostic radiologists all review the same x-ray scans for all patients in a clinical sample (cf. Wagner, Beam & Beiden, 2004, for an example of a study using MRMC). Using this approach the difference in sensitivity, across the participating clinicians, can be taken into account when accessing diagnostic performance. Thus, by evaluating information gathered from a carefully selected sample, conclusions on the effectiveness of a relevant test or imaging technique can be generalized across the population of intended users.

The “Localization Receiver Operating Characteristic” (LROC) and the “Free-response Receiver Operating Characteristic” (FROC) are techniques that augment standard ROC methods by adding a localization element to the search for targets.

For example, in a study that follows an LROC design, the observer (e.g. a clinician screening a mammogram for a tumour) in addition to selecting an image (e.g. a mammogram) as containing a target of interest (e.g. a cancerous lesion), then has to also identify the location of the suspected anomaly (Metz, 2006). An obvious limitation of this technique is that it is only suitable in cases where images may potentially contain only a single target of interest.

Similarly, in the FROC approach, the observer is again able to qualify where the perceived target or anomaly is in a hypothetical medical scan. However, by contrast, in this method the observer is free to select as many locations of interest as he or she deems appropriate (Metz, 2006). These are tagged and rated as to their perceived probability of containing a target of interest. Again returning to the previous example of the clinician examining a mammogram, in this instance there might be several suspected lesions locations, all of which can be identified and categorized accordingly.

Thus the difference between the LROC and the FROC is that the latter supports multiple potential target locations, within the same image. Consequently, these two techniques differ in the way the data are recorded and treated in the subsequent analysis.

2.3.3 Present Applications

There are numerous kinds of SDT experiments, and both single and dual interval experiments were considered. Additionally, as seen in the preceding discussion, alternative ROC methods could be applicable. However, it was not practical to implement these more complicated approaches in the present research. Consequently, in the interest of straightforwardness the basic ROC approach was selected for the purposes of this project. Hence, the discussion here will primarily focus on simple detection and discrimination aspects of SDT, as well as expand on confidence rating methods.

The Yes-No experiment is the classic example researchers bring to mind when considering SDT experiments. The observer is confronted by a single stimulus at a time; hence it is referred to as a single-interval design. He or she then has to decide if the target is present or absent by replying "Yes" or "No", which is where the common name for this experiment originated. In this instance it was decided that a variation of the basic Yes-No design, called the rating experiment was the most appropriate implementation. Adding a rating component to the experiment gives the opportunity to plot more detailed ROC curves, which allows the experimenter to compute measures of both sensitivity and bias, thus permitting a better evaluation of performance.

The potential outcomes in this type of experiment will now be examined. In the case of a signal trial, where the target is present, and if the observer detects it, then this instance is considered a "Hit" (H). On the other hand, a 'False Alarm" (F) is the instance when a target is erroneously selected as being present during a noise trial. Similarly, if the target is correctly identified to be absent during a noise trial this instance is defined as a correct rejection. However if the observer fails to detect the target during a signal trial this is defined as a miss. An illustration of these different contingencies is presented in the table below (see Table 2), which has been created in order to represent the potential outcomes of this research project.

Table 2. *The stimulus-response contingency table used in Yes-No experiments***.**

Yes-No Design	Responses		
Stimulus Class	Image Contains a Threat	Image Does not Contain a Threat	N
<Yes Threat Image>	Hit (True Positive)	Miss (False Negative)	
<No Threat Image>	False Alarm (False Positive)	Correct Rejection (True Negative)	

Note. H= ("Image Contains a Threat"| <Threat Image>), F= ("Image Contains Threat"| <No Threat Image>), N= number of responses.

In 2AFC experiments two stimuli are typically presented, either simultaneously or sequentially, hence this is why it is sometimes referred to as a two-interval design. The observer has to choose which interval contained the signal, or in other words discriminate between the two stimuli (cf. Katkov, Tsodyks, & Sagi, 2006, for an example of a 2AFC design). Since participants are compelled, by the design of the experiment, to choose either one or the other alternative presented it is regarded as a forced-choice procedure. Other designs, such as the mAFC, exist, where the observer has a choice of more than two stimuli (cf. Eckstein, Ahumada, & Watson, 1997, for an example of a mAFC design), however the discussion here will focus only on the two-interval variant. The advantage of using 2AFC procedures is that they tend to discourage bias and are also more sensitive to smaller inter-stimulus differences (Macmillan & Creelman, p179, 2005). As in the case of single interval designs, adding a rating component to the 2AFC experiment yields additional chances for data analysis, such as the option to conduct an ROC analysis if this is deemed appropriate.

In order to analyze 2AFC data the differences, in this design, of the definitions of hits and false alarms must first be examined. More specifically, the case where the design calls for sequential presentation of the stimuli. A "Hit" (H) is defined as the instance when the first stimulus presented is deemed to contain the target, when the order of presentation is <signal first, noise second>. Thus a hit is correctly selecting the signal stimulus as the one containing the target, when it is presented first. On the other hand, a "False Alarm" (F) is defined as the instance when the first stimulus presented is deemed to contain the target, when the order of presentation is <noise first, signal second>. Thus a false alarm is erroneously selecting the noise stimulus as the one containing the target when it is presented first.

The application of this design to the current project is illustrated in the table below (see Table 3) with the appropriate contingencies as expected in the present experiment.

Table 3. *The stimulus-response contingency table used in 2AFC experiments*.

2AFC Design	Responses		
Stimulus Sequences	1st Image Contains a Threat	2nd Image Contains a Threat	N
<Yes Threat, No Threat>	H	1-H	
<No Threat, Yes Threat>	F	1-F	

Note. H= P ("1st Image Contains Threat"| <Threat, No Threat>); F= P ("1st Image Contains Threat"| <No Threat, Threat>), N= number of responses.

The rationale underlying the implementation of 2AFC designs is that they are well suited to instances where the signal is weak. Also, a 2AFC design better protects against observer bias. For instance, in situations where you have presentations of stimulus pairs it is possible for the participant to exhibit a recency effect, meaning that in an ambiguous situation the second image (and temporally closer option) might be preferred as the one containing a threat. This tendency is clearly counterbalanced in the design, because in the 2AFC experiment the presentation order of the stimulus pairs is split evenly in terms of signal and noise.

Additionally, because of the way the model is constructed it assumes equal variances of the signal and noise distributions (Macmillan & Creelman, 2005) thus the d' is always considered a reliable measure. Finally, percent correct in this design is synonymous to the area under the curve (A_z) making it a reliable performance index, although as previously stated some researchers maintain that under certain conditions percent correct is a reliable index in yes-no experiments as well (Balakrishnan, 2003).

2.4 Making Sense of a Complex Environment

But before any investigative approach can be implemented, an understanding of the factors that regulate performance in the security screening task is required. Ultimately this can help recognize the reasons for an operator's failure at this task, so possible methods of remediation can be considered. A large body of research exists in this field, as this investigation has recently gathered increased momentum as a result of the unfortunate events of the past decade. Topics of interest include the perceptual learning processes that screeners undergo during their training, efficient strategies of visual search, how objects are recognized and how to address the problem arising from the fact that the actual targets (threatening objects) occur very infrequently.

2.4.1 Perceptual Learning and Object Recognition

The issue of perceptual learning is a complex one and requires precise and thorough methods to investigate. In brief, perceptual learning refers to the enhancement of complex perceptual skills as a result of training (Sowden, Davies, & Roling, 2000). As this is an extensive field, for the purposes of this research, the focus will be on the particulars of perceptual learning during visual inspection tasks.

Roling et al. (1997) examined the perceptual learning processes that take place during inspection tasks. Specifically by conducting three experiments they looked into the detection of low contrast features set in a noisy background. They noted that both sensory learning, which is considered a low level process, and conceptual learning, a higher level process, contributed to the participant's performance in the tasks. Furthermore, they suggest that, with better training, detection abilities in inspection tasks (e.g. examining medical x-ray images) can transfer from easier to harder tasks.

In more recent experimental work they went further and directly investigated the role of perceptual learning for the detection of features in medical x-ray images (Sowden, Davies, & Roling, 2000). They also looked into whether this process plays any functional role in improving visual sensitivity. Their findings suggest that visual analysis sensitivity, for detecting critical abnormalities in x-ray images, is enhanced by experience. Furthermore, they propose that persistent feedback training with exposure to as many x-ray images, with clinically-relevant structural abnormalities, as possible should be a component of medical x-ray teaching programs.

The recommendation of emphasizing constant explicit training is also suggested by a common sense approach to the problem. However, at this point it would be beneficial to the discussion to describe some types of implicit perceptual learning. Contextual cueing is an implicit memory effect that occurs in tasks with repeated displays (Chun & Jiang, 1998). Through repeated presentation of the stimuli global patterns of the scenes are learned by the observers and help automatically guide their attention to the target. This generally leads to rapid and accurate responses.

Peterson and Kramer (2001) investigated the mechanics of contextual cueing by monitoring eye movements. They report that recognition does not necessarily occur on every trial and sometimes only affects the outcome of the search later in the process. In the cases that recognition took place the participants' attention did not simply turn to the general location but was guided straight onto the target. However, Kunar, Flusberg and Wolfe (2006) caution that while global features of a scene are essential in contextual cueing their role is somewhat limited. These global properties only function as guiding features if enough exposure time is provided and if they are explicitly associated with the location of the target.

There is however consensus in the literature that contextual cueing leads to a general speeding up of reaction times (Chun & Jiang, 1998; Kunar, Flusberg, & Wolfe, 2006). While due to the extreme heterogeneity between different x-ray scans (McCarley & Carruth, 2004) it is not really an issue in the field, this effect has to be taken into account when conducting visual experiments that incorporate repetitions of trial stimuli in their design.

The problem of assessing these mechanisms is further complicated by dissension in the field on their perceptual order. For instance Peterson and Gibson (1994) argue that figure-ground segmentation is not a necessary prerequisite for object recognition, but instead early stage object recognition processes can inform the perceptual system into making judgements about figure-ground relationships.

One could relate this finding with the concept of uniform connectedness (UC), a principle of perceptual organization (Palmer & Rock, 1994). Uniform connectedness states that closed regions of homogeneous properties (e.g. lightness, chromatic colour or texture) tend to be initially interpreted by the perceptual system as single units.

The fact that this perceptual grouping occurs irrespective of the depth field - the fraction of the scene that is in focus and thus adequately sharp - (Palmer & Rock, 1994), is of particular importance as the very nature of static x-ray images preclude the presence of any depth cues. Hence this initial perceptual grouping can lead to false impressions when examining x-ray images and lead to either false positives or false negatives depending on the different greyscale gradients in each case.

Pizlo (2000) suggests a new way of thinking about perceptual interpretation and postulates that the perception of depth, motion, colour and lightness, object shape as well as figure-ground segregation requires a-priori constraints (e.g. minimum variance of angles, planarity, etc) because the information available in the retinal image is inherently ambiguous and could lead to countless alternate interpretations. It then follows that 3-dimensional shape perception involves a-priori constraints even when depth cues are strong. The stated implication of adopting this approach would be a re-think of the way the interpretation of the physical world, in terms of a stimulus-specific hierarchy of processing stages, is conceptualized. This is of interest as if this opinion is accepted, then there are consequences for how threatening targets are identified in x-ray images.

At its most basic level object recognition relies on correctly fusing together the contours of the object, and then accessing and comparing the perceived object to mental templates. Again as this is a diverse research area the discussion will focus on a small section that is considered as the most relevant to this research.

Hayward (1998) investigated the use of outline shape in recognizing objects. What is of particular interest is the technique used in these experiments. Specifically participants were presented with shaded images and silhouettes of the objects they were called upon to identify. The general finding was that outline shape is sufficient for object recognition, but this involves certain caveats. Specifically he notes that similarities in outline shape of the presented stimuli predicted recognition performance. The relevance to the present research is readily apparent as perceptually x-ray images are shadows of the physical objects being scanned (Evans, 2002).

2.4.2 Visual Search

Visual search is a perceptual task that requires the observer to actively scan the environment for a particular feature or object (the target) among other features or objects (the distractors) while maintaining attention and efficiency (Wolfe, 1998). The visual search paradigm is an integral part in the design of the present experiments in as far as the participants looked for a target item among numerous distractor items, and the dependent variables were response time (RT) and some measure of accuracy (Wolfe, 1998).

The psychophysics of visual search has been systematically investigated, and different theories regarding the underlying mechanisms have been postulated. Notably, Palmer, Verghese and Pavel (2000) have proposed a signal detection framework for empirically evaluating visual search. This way of analysing visual search has several benefits as it allows the consideration of visual search in simple terms as a detection and discrimination problem and avoid complex serial versus parallel processing arguments. However, while an SDT conceptualisation of visual search can be empirically useful, especially while designing experiments, it is also useful to delve deeper into the theoretical underpinnings of this process.

There are two major types of models trying to explain visual search: serial and parallel (Wolfe, 1998; Horrowitz & Wolfe, 1998). Serial models state that each item is processed individually. If it is classified as the target the search halts, if it is classified as a distractor the search continues to next item and so forth. This model also assumes that there is an inhibitory mechanism that prevents attention from refocusing on previously examined items. Parallel models on the other hand maintain that each item is evaluated in parallel and the judgement on the identity of each item develops gradually over the course of the search. A decision is reached when the target is found or all the items are deemed as distractors. Thus, both models assume that visual search is memory driven and that the visual system needs to accumulate information about the scene in order to reach a decision (Horrowitz & Wolfe, 1998).

However, Horowitz and Wolfe (1998) postulate that visual search has no memory. By constantly shifting the contents of the scene (by employing a brief flicker or an intervening object) they concluded that there is only a small amount of visual information integration across saccadic eye movements. What they mean by this is that the visual system appears not to accumulate information about object identity over time during the search process. If this position is accepted, then it becomes imperative to make the visual search process as efficient as possible.

An interesting piece of evidence comes from research by Park et al. (2004) into the effects of chemically induced temporary amnesia on visual search. They found a failure of both explicit as well as implicit memory functioning. Specifically, contextual cueing (Chun & Jiang, 1998) did not take place during repeated trials. However the general speed up of trial times, which typically occurs with practice, was unaffected. This indicates that procedural learning took place; hence acquisition of skills was not compromised. Of course their aim was to further scientific understanding on the processes that underlie amnesia. However they provide compelling evidence of how a truly "amnesic" visual search might be conducted and the implications of such findings for the processes of visual search itself cannot be ignored.

The efficiency of a visual search exercise can depend on the search pattern, evident from the eye movements necessary to scan the image. Najemnik and Geisler (2005) describe what they regard as the optimal eye movements in visual search based on an ideal (Bayesian) observer (or in this case searcher) analysis. In their model an ideal searcher is precisely aware of its own visual system (e.g. high fovea sensitivity/low peripheral acuity) and is able to utilize precise knowledge about the makeup of the scenes, in which the target is positioned, in order to perform eye movements that gather the most information about the target's location. Furthermore, they maintain that there is little benefit from perfect integration across fixations, but rather the true benefit is in the efficient processing of information on each fixation. Apparently, humans must have evolved to take advantage of this mechanism in order to achieve efficient eye movement strategies with minimal neural resources devoted to memory. This last point concurs with earlier findings (Horowitz & Wolfe, 1998) on the nature of memory utilization during visual search.

Peripheral information is also important in visual search. Geisler, Perry and Najemnik (2006) investigated the role of peripheral information in visual search by employing gaze-contingent displays. They note that humans are very efficient at visual search in complex naturalistic backgrounds and that many human eye movement statistics parallel those of an ideal (Bayesian) observer. This suggests than in order to perform so well, their participants performed efficient parallel processing across the search area on each fixation, as well as selected the fixation locations with high efficiency. Lastly a good searcher must exhibit an inhibition of return, as refocusing on an already searched location wastes time and limits the efficiency of the search. This last part is of paramount importance, as their results also indicate that memory for image details and ability to integrate across fixations adds relatively little to search performance. They do however also caution that in their research the target was always present, and theorize that in cases where the target is absent (noise) or the signal is very weak the efficiency of the search may be reduced due to increased memory demands.

The efficiency of the visual search, when the target is in the periphery, is also of importance because humans generally have a tendency to focus on the middle of the display and this affects the chances of detection, especially when dealing with a weak signal. Thus research into efficient search patterns when dealing with peripheral targets can potentially guide the development of more effective training regimes for screeners.

2.4.3 Visual Masking and Crowding

Visual search is further complicated when the other objects in a complex display interact with the target by either partially or totally occluding it. Different levels of target occlusion based on amount of overlap, contrast and spatial frequency of the distractor objects can produce distinct perceptual effects.

In general, visual masking is the reduction or elimination of the perceived visibility of a stimulus (target), by the presence of a second stimulus (mask) (Olzak & Thomas, p10, 1986; Howard, p333, 2002). Ordinary masking distorts both feature detection and identification and does this most effectively when the mask has similar spatial frequency, orientation, and location in the display as the target signal (Pelli, Palomares, & Majaj, 2004). The mask prevents processing of the target and thus renders it invisible to the observer (Kim & Blake, 2005). The specific type of masking, of concern here, is long duration masking caused by occlusion of the target object by other objects in the x-ray display.

While spatial-frequency and contrast properties of ordinary masking and crowding are up to a point similar (Chung, Levi, & Legge, 2001), and they do indeed share some processes (Huckauf & Heller, 2004) there are functional differences at a deeper perceptual level. As aforementioned, the latter has negative effects on both detection and identification, whereas the former affects only identification, and not detection (Pelli, Palomares, & Majaj, 2004).

Crowding is principally a failure of object recognition. An object is “crowded” when background clutter distorts its appearance, thus making it impossible to identify it (Pelli et al., 2007). Crowding is unlike ordinary masking, where there is an absence of the masked object’s features. Crowding frustrates object recognition by preventing feature integration (Pelli, Palomares, & Majaj, 2004). The different features of the target are visible, but the mask prevents the perceptual system from recognising them as an object (Kim & Blake, 2005). This mechanism is of particular importance because the KDEX technique (presented earlier in the text) potentially enables the screener to overcome these problems.

Additionally research has shown that crowding degrades saccadic search performance (Vlaskamp & Hooge, 2006) and thus reduces the efficiency of visual search. Bex and Dakin (2005) suggest that crowding probably arises from higher-level integration processes in the visual system, but also depends on lower-level factors, such as the configuration of the components of a scene in the retinal image. There is however consensus in the literature that crowding increases with the eccentricity of a scene (Chung, Levi, & Legge, 2001; Bex & Dakin, 2005), and Gale, Done and Frank (2001) demonstrated that it causes category-specific deficits in the perception of pictorial stimuli.

Pelli and Tillman (2008) report the existence of an "uncrowded window" in visual perception. This postulate relates crowding to the broader visual search paradigm, because the implication is that even if the target is not occluded it can still be missed if it is not located within the uncrowded window of the observer, or in other words if he or she does not fixate upon it.

As will be seen later on, this last finding is relevant for target detection in x-ray images because the nature of the stimuli themselves can give rise to these kinds of crowding effects. This could result in a potentially threatening object being missed. At the same time, colour differences due to materials differences may well ameliorate effects of crowding.

2.4.4 Target Identification in X-Ray Imagery

The failure of airport security screeners to identify threats during field training simulations is often reported in the news media (CNN, 2002). A wide body of research has formed around the problem of error misses in x-ray security screening and many different approaches have been proposed as to the underlying mechanisms as well as methods of amelioration. As seen earlier, the background theory describing the security screening task combines elements of perceptual organization, visual search, object recognition as well as signal detection and vigilance. The focus will now be on how these different theories combine to describe performance and lack thereof in x-ray screening tasks.

Perhaps the closest research area to the present topic is the field of medical imaging. Studies on medical x-ray image recognition (Myles-Worsley, Johnston, & Simons, 1988) indicate that radiological experience had a selective effect on performance, which was concentrated on memory for severe abnormalities in chest x-ray images. Unexpectedly this was to the detriment of normal (non-pathological) x-ray images. So in this case expertise selectively biased the participants to develop expertise in detecting abnormalities. The general conclusion was than in order to perform well radiologists had to know both the normal appearance of an x-ray image as well as the clinically relevant features that signal pathology.

Additionally, the medical imaging literature also informs that the saliency (or conspicuousness) of the target of interest (e.g. a nodule) is also important for detection; sometimes even to the expense of its relative size. Thus a smaller, but more salient nodule will be detected more readily, than one that may be larger, but is however less conspicuous (Krupinski et al, 2003). This is an interesting finding with relevant implication for the security screening process. For example, an improvised explosive device (IED) may be less conspicuous than a small pistol, even though the former may be larger.

Manning, Gale and Krupinski (2005), comment on ways to improve diagnostic performance in medical x-ray images. Their recommendations are threefold and include first making improvements in both the quality of the information, as well as its presentation; second, improving the taxonomy of critical features; and third improving the detecting ability of radiologists. The emphasis is on better performance assessment measures and subsequently better training of practitioners.

Among their other recommendations is that caution needs to be taken when selecting the monitor upon which medical images are displayed, as differences in specifications can affect the speed and accuracy of the image interpretation. This is also supported by the findings of Krupinski and Roehrig (2002) in a study which compared performance on locating pulmonary nodules in chest x-ray images, across 3 different CRT display types (2 monochrome, 1 colour). Specifically, they found significant differences in accuracy between the 3 displays. Interestingly enough, diagnostic performance on that task was worst in the case of the colour display.

Based on these and other relevant findings, similar considerations could then be applied to the process of interpretation of security x-ray images. Radiological diagnostic work using x-ray images far predates their use in airport security screening. Consequently research into the different factors affecting performance in radiology has been on going for decades (cf. Manning, Gale & Krupinski, 2005). It is thus imperative to examine the common ground, both in terms of research methodology and practical application, between the two areas. Additionally, the reader is encouraged to visit the MIPS website (http://home.comcast.net/~eakmips/), as it is an excellent source of further reading into the wealth of relevant research within the medical imaging area.

If the two tasks are directly compared it can be seen that there are numerous similarities between security screening and medical radiology, such as looking for a usually weak signal in a complex background. After all often in preventive medicine the aim is to identify a tumour before it becomes too large to tackle, because a tumour detected when it is too late to operate would be as diagnostically useful as discovering a bomb after it has gone off. Similarly a purposeful attacker will undoubtedly attempt different methods of concealing whatever contraband he or she tries to smuggle on board an aircraft.

There are however noteworthy differences however between medical images and airport security images. For instance, despite sexual dimorphism and general structural differences because of differences in build and age, human bodies offer a fairly consistent background in x-ray images (McCarly & Carruth, 2004). After all, the normal structural appearance of a healthy human body is generally well known by clinicians. On the other hand there can be a near incalculable range of combinations for how the contents of bags can appear on an x-ray scan. This negates any sort of contextual cueing (Chun & Jiang, 1998) from taking place and help guide the screener to the target. But as was previously discussed, contextual cuing can occur in a laboratory setting depending on the design of the experiment.

Interesting research, specifically into the overlap between the fields of medical and security x-ray image inspection, has also been conducted by Gale et al. (2000). By examining the factors involved in airport security screening they proposed applying a clinical model to the problem of missed targets in security x-ray images. Specifically, in their study they compared and contrasted observer performance, of both expert and inexperienced participants, while searching for improvised explosive devices (IEDs) hidden inside luggage. ROC measures of detection performance, between the two different groups of observers, were augmented by the recording of eye tracking data. By studying this data, they concluded that a medical conceptual model is applicable to the description of the process of target detection in airport security x-ray images. Such a model offers a way of describing the cognitive process during a security x-ray image inspection, and identifying and categorising any miss errors. Consequently this allows for effective strategies to be devised for reducing these failures of target detection.

More recently, Fiore, Scielzo and Jentsch (2004) examined the perceptual learning processes during security screening. More specifically they looked into the contribution of varying degrees of occlusion by masking objects and general background clutter to target identification in x-ray images. They accomplished this by conducting simulated x-ray screening experiments using non-screeners. Their main findings were that the learning that took place was influenced by each participant's individual characteristics, and that it is crucial for evaluating different training regimes to use as variable a target set as possible. They also note that being able to mentally manipulate the objects in a complex scene is just as important to recognition as detecting their presence. Their findings are useful, with regards to the potential composition of the present stimuli. Most importantly it is felt that the introduction of kinetic stimuli would remove this need for the screeners to mentally rotate the objects, as this will be done for them by the proposed new system. This will reduce the mental strain of the task and allow them to focus their attention on target detection.

Also working with inexperienced participants Liu et al. (2006) examined their threat detection performance while looking for knives, guns and IEDs in a visual screening task. They found that the familiarity of different threat items significantly influenced the rate at which they were detected by the participants. Also, by tracking the participants' eye movements they discovered that bag complexity and the conspicuousness of the different threatening items influenced their attention to them. The obvious implication here is that previous knowledge of threatening objects influences their detection and recognition in a subsequent inspection task. Furthermore, they comment that, given a short period of familiarization to the way threatening objects appear in x-ray images, novice participants were able to perform better than chance when they subsequently examined luggage scans for other threatening items.

Other researchers (McCarley et al., 2004) also conducted simulated screening tests with inexperienced participants. They report that participants' sensitivity and response times improved reliably as a result of practice. Additionally they tracked their eye movements during the task, and the data revealed that sensitivity increases were produced entirely by changes in their ability to recognize the target objects and not by changes in the effectiveness of visual scanning. Moreover, when they introduced unfamiliar targets performance degraded, which led them to conclude that the recognition skills acquired during the experiment were partially stimulus-specific. This is consistent with earlier research findings in the area of perceptual learning. For instance, Furmanski and Engel (2000) measured the transfer of perceptual learning in object recognition and found it to be specific to the practiced objects. This was irrespective of changes in image size and independent of the experimental context.

With respect to occulomotor scanning, in a further paper McCarley and Carruth (2004) discuss how eye tracking results can help explain why perceptual lapses occur during searches for targets in complicated backgrounds. This knowledge of the limitations of observers in inspection tasks allows the anticipation of the conditions that frustrate target identification and stress the need to evaluate screener training in order to ameliorate them. In fact they point out that better imaging techniques, incorporating for example 3D displays, should help facilitate perceptual organization by sensitivity by making the targets more salient. Also the issue of effective visual search strategies comes to mind, as the two areas are interconnected, and this topic will have to be addressed later in the text.

Taken together these findings stress the fundamental role of screener training should be to expose them to as numerous and distinct sets of targets as possible so as to increase their knowledge base.

In a large study Schwaninger, Hardmeier and Hoffer (2005) compared the differences of screeners to non-screeners by examining the influence of image-based and knowledge-based factors for detecting threat items in hand luggage.

They note that image-based factors such as bag complexity, viewpoint, and superposition of the target by distractor objects (i.e. masking or crowding) can substantially affect detection performance. They found that experts (screeners) were moderately better than novices (non-screeners) in detection performance, with regards to image-base factors. However large within-group differences were found for both screeners as well as novices, suggesting that even experts tend to apply different criteria to the detection task.

As expected, screeners faired considerably better with regards to knowledge-based factors, such as knowing which items are prohibited and how they appear in x-ray images of hand luggage. This illustrates the fundamental difference between experts and novices in the inspection task, the former having better visual knowledge based on training and experience.

Overall, the study's results support the notion that performance in x-ray image detection tasks depends both on visual abilities, which are necessary in order to make sense of the scene, as well as training and overall experience. The latter two qualities enable the observer(s) to know which objects are targets and what they look like in x-ray images.

Research in comparing performance differences between screeners and non-screeners was also conducted by Liu, Gale and Song (2007). In their experiment they recorded observers' eye movement patterns while looking for threatening targets in x-ray images. They found that the specific advantage of professional screeners, when compared to non-screeners, was their speed and accuracy in detecting the threats present in the luggage scans. Furthermore, based on data collected on eye position, they concluded that screeners fixated on the target areas faster than non-screeners. Also, once they fixated on the relevant target areas they achieved significantly higher hit rates than non-screeners. Thus the screeners' experience with x-ray security screening enabled them to conduct a visual search that was both more efficient, and more effective than their non-screener counterparts. This is consistent with, and also complements, earlier findings (Schwaninger, Hardmeier & Hoffer, 2005).

Taken together, these findings are actually quite relevant since novice volunteers (typically university students) make up the overwhelming majority of participants in screening experiments to date. Hence, in order to meaningfully simulate the screening task in the laboratory these differences need to be taken into account, and the task moulded accordingly. Additionally, as much training, as time constraints permit, needs to be given to the novice participants.

With regards to individual differences in threat detection, Washburn, Taglialatela, Rice and Smith (2004) examined the effects of time passage on participants' performance in an x-ray screening task. Their participants completed a 25 minute long security screening task that asked them to locate guns, knives and explosives. They report that the effects of having to maintain sustained attention over the duration of the experiment varied across participants. The obvious implication is that a cognitive assessment of applicants' ability to maintain vigilance would be a beneficial component of the selection process for professional screeners. A less obvious point however is that adopting new technologies and procedures that can reduce the complexity of the screening task can also facilitate performance in x-ray screening tasks.

Another issue confounding screeners is the variability between different threatening object categories (e.g. guns, knives etc) and their near infinite members. To address the effects of this target diversity on screening task performance Meneer et al. (2004) conducted a series of experiments comparing and contrasting single and dual target searches. They found that performing a search for two targets was less efficient than performing two separate searches, one for each target. The implications are of course clear, especially considering that the screening task involves multiple target category searches (e.g. handguns, knives, drugs, IEDs, and more recently liquids).

Security screening can also be conceptualized as a categorization task. Smith, Redford, Washburn and Tagliatela (2005) evaluated the effects of visual complexity on the categorization processes taking place during an x-ray inspection task. By conducting simulated screening tasks they observed that repeated training with a certain set of stimuli led to increased performance improvements However, the gain evaporated once the stimuli were changed. This they maintain provides evidence that there were specific token effects influencing the participants' performance, rather than general category factors.

Thus they caution that sampling targets from a relatively restricted library generates repetition. This repetition of target stimuli leads to the development of familiarity with them; which invariably causes the participants to become dependent on recognizing familiar targets that they practiced with. They engage in this kind "cognitive short cut" instead of applying category-general knowledge to the task at hand. This finding has implications on how screener training is conducted, which will be discussed later in the text.

One interesting approach to boosting performance in object recognition is to enlist cue recruitment. Haijiang, Saunders, Stone and Backus (2006) experimented with perceptual learning aided by cue recruitment using Pavlovian (classical) conditioning. They discovered that introducing a new signal in this context led to the recruitment of this new signal, by the visual system, as a cue for the construction of visual appearance. Thus, it is suggested that cue recruitment can be introduced into a detection task and possibly aid discrimination performance.

However, there are additional problems complicating target detection in x-ray images.

2.4.5 The Problem of Infrequent Targets

Another issue that needs to be addressed is the mismatch between target prevalence in an experimental setting and the actual environment. This is further complicated by the knowledge that rare items often tend to be missed in visual searches (Wolfe, Horowitz, & Kenner, 2005). This creates a problem for performance in inspection tasks.

In a recent lengthy paper Wolfe et al. (2007) proposed a way to address frequent miss errors, in screening tasks where the target occurs at a very low probability. Their proposed solution is to populate the screening task with essentially "dummy" targets and thus artificially raise the target prevalence with a view to raise the alertness of the screeners, which they claim suffers in low probability searches. Their aims appear to be in part to produce a criterion shift, which is a change in the observers' detection decision criteria (cf. Singer & Wixted, 2006, for a discussion on factors that influence criterion shifts) and thus reduce miss errors.

However, almost immediately Fleck and Mitroff (2007) contradicted this view and instead proposed that missing these targets was a mistake that observers often found themselves making because of being accustomed to saying no in a visual search task. Their solution to the problem of miss errors in low probability searches is a correctable search paradigm, where observers can go back and change their answers if they feel they made a mistake.

However, it is likely that both these procedures reveal more about how the laboratory conditions affect performance, rather than actual behaviour. Thus, the latter is not directly relevant, because if screeners in the field have any reservations about an x-ray image of a particular bag they can very simply open it up in order to conduct a visual search.

As an interesting conclusion, further review of the literature suggests that video gamers may have superior detecting abilities because of their gaming experience. Green and Bavelier (2007) reported that playing action video games improves the spatial resolution of visual processing. Furthermore, Fleck and Mitroff (2008) also reported that individuals with extensive videogame experience perform much more accurately on such rare target visual searches. These types of results demonstrate an innovative approach to improving performance in visual inspection tasks.

2.4.6 Summary

But now it is time to consider the implications of all of these research findings on this project. Realizing that while it is necessary to keep the task as relevant to the real world conditions as possible, at the same time conducting meaningful experimental work was a priority. Hence, the author chose to maintain the classic SDT 1:1 signal to noise trial ratio in this research. In fact, if this research tried to strictly adhere to the proportion of actual threatening objects that are confiscated in airports, out of the literally millions of luggage items that are inspected each year only a relative small proportion are actually found to contain serious contraband (CNN, 2003). Even further, guns are such an infrequent occurrence that in order to maintain a relatively strict environmental prevalence there would have to only be one trial where the target is a gun for every thousand trials where the target is a knife.

It is the nature of the screening task that certain factors may influence performance. For instance, the pressure imposed on the scanner operator by being situated in a busy airport with a queue of people waiting to go through security clearance may negatively affect them. It seems obvious that given the potential outcome of miss errors the operators should be encourage to err on the side of caution when selecting targets, even if this means increasing false positives (false alarms). Yet there does not seem to be a general consensus in the security community on how to best achieve that.

The current emphasis is on constant training of the operators by digitally inserting targets in the field to maintain alertness. This is achieved by a process known as "Threat Image Projection" or TIP. However, the problem with this method of testing and evaluating TSA screeners is that depending on the size of the set of possible objects inserted implicit memory mechanisms can confound the results (cf. Smith, Redford, Washburn, & Tagliatela, 2005). Repetition priming, the effect that prior exposure to a particular stimulus has on a later perception or on a later decision about it (Ratcliff, & McKoon, 1996) can positively affect the results. In other words if the number of objects from the database is not extensive enough, then the possibility of repeating objects increases and screeners may be detecting the target because they encountered it before, thus providing a false sense of security. This of course can be counteracted by increasing the number and variability of possible inserted targets to make it statistically unlikely that the same or similar objects are repeated.

It is an obvious conclusion from the many studies reviewed here that because of the cognitive capabilities and limitations of the human perceptual system, with regards to perceptual organization and visual search, adopting any new technology that can increase the saliency of the x-ray image, and thus reduce the need for top-down processing (McCarley & Carruth, 2004) will lead to a performance improvement for screeners. This expectation alone more than justified the effort put into pursuing the present investigative course.

Ultimately the best way to tackle airport security is a combination of new technologies and refinement of training procedures to capitalise on the lessons of cognitive psychology research (cf. Herrmann, 1996).

2.5 Human Factors Considerations

In general, “human factors” is an umbrella term that covers the variety of disciplines that deal with the interaction of humans with their technology. As with SDT and ROC analysis, human factors research had its origins as a discipline in military research during World War II, and applications to civilian tasks followed suit (McCormick & Sanders, 1982).

Simply put it can be referred to as a systematic multidisciplinary approach of designing for human use. More specifically, human factors deal with “the consideration of human characteristics, expectations and behaviours in the design of the things people use in their work and everyday lives and of the environments in which they work or live” (McCormick & Sanders, 1982).

This project was a laboratory based performance evaluation of the KDEX imaging technology. However, the actual simulated screening experiments were conceived in their entirety to be conducted on a computer, as was part of the training the participants received. Accordingly, some considerations had to be incorporated into the design of the experimental interface to ensure the reliability and validity of the results.

Since, these experiments incorporated features of interactive computer systems, elements of both human computer interaction (HCI) and computer assisted learning (CAL) had to be applied in order to design the user interface and streamline the entire process. Hence, how the prospective user (in this case experimental participant) would interact with the computer, first had to be examined, and then steps taken to try and facilitate this process.

The objectives were to make the task as efficient as possible, while at the same time ensure that the procedure made it comfortable for the participants to partake in these experiments. The approach was to assess the expected capabilities and limitations of the potential study participants (novice volunteers) and optimize the task to fit best the user.

By taking all these points into account, the conclusion was that the user interface had to be easy to navigate, largely self explanatory, and comprehensible even to users with the most basic of computer knowledge. This would reduce the complexity of the task and, after a reasonable familiarization period, allow the users to focus their cognitive abilities on the actual task of finding potential targets in the different x-ray displays.

2.6 Closing Thoughts

Current day airport security procedures require a human operator to monitor the x-ray scan of the carry-on bags and watch out for and report any suspicious objects. As an alternative approach, it has been proposed that there is a need to establish automated threat detection systems. For example, Harris (2002) defined an ideal threat detection system as an automated one that could identify a threat in 6 seconds, and hand over control of the situation resolution to a human supervisor.

One example of such an approach is implementing automatic object recognition using edge detection (cf. Green, Blumenstein, Muthukkumarasamy & Jo, 2005; Oertel & Bock, 2006). However, while this research has merit, these types of automated technologies are as yet in their infancy. Consequently, they are not capable of replacing human screeners in the foreseeable future. Furthermore these approaches adopt a technology only view to security research; this is in contrast to the more desirable human operator screening approach, which can be improved through cognitive science (Fiore, Rubinstein & Jentsch, 2004).

In order to substantially improve airport security, more research is needed on better screener training regimes (cf. Gale et al, 2005; Michel et al., 2007), as well as better technologies to assist the human operators during the security screening task (cf. Fiore, Jentsch & Rubinstein 2004). Here on offer is a system that will readily work within the existing security framework, and potentially achieve a higher level of accuracy (Evans, 2003; Vassiliades et al., 2008).

3 Chapter Three: Experimental Work

This chapter contains a comprehensive description of the experiments conducted in the course of this research project. To establish the most suitable research design the first step was to plan and execute a series of pilot experiments. This was done in order to formulate and test different approaches, establish the appropriate procedures, calculate the time commitment required out of potential participants, as well as identify all the smaller issues that accompany this type of experimental work. The pilot study was then followed by 5 main experiments.

However, it should be noted that the development of the x-ray scanner was ongoing throughout this programme of research. Consequently the quality and availability of stimuli were affected especially during the early work. Specifically, for experiments 1 to 3, as well as the pilot, the KM stimuli were limited to an angle span of -7°/+8° (16 frames). Additionally, it was not possible to produce good quality pseudo-colour x-ray images. Thus, these experiments were conducted using the available greyscale stimuli. Furthermore, the equipment issues meant that only a limited number of x-ray images could be collected. Consequently, only images of bags containing knives as a target were produced.

In the intervening period, the experimental scanner had to undergo a series of technical upgrades. Henceforth, the quality of the KM images in experiments 4 and 5 was substantially improved. The angle span of the stimuli was expanded to -20°/+20° (41 frames), and good quality pseudo-colour images could now be collected. The range of threatening items was also increased, in experiment 4, to include handguns, scissors and grenades, as well as knives.

3.1 Pilot Study

The pilot study was exploratory in nature. As discussed in the previous chapter, signal detection theory (Green & Swets, 1966) was selected early on as the theoretical framework for studying the present problem. After some thought it was concluded that before continuing it would be beneficial to empirically ascertain which type of SDT experiment would be most appropriate to use in this instance.

In the end it was decided to conduct two separate pilot experiments using both a Yes-No single interval detection design, and a two-alternative forced choice (2AFC) discrimination design. To conduct these experiments unpaid volunteers were used, recruited from the University student and staff community. The design and results of each experiment will be briefly reviewed, and an explanation given as to the rationale behind settling on a yes-no (single trial) design. The methodology discussion will focus on the procedure and experimental design as the stimuli and apparatus will be described in detail in the following sections (main experiments).

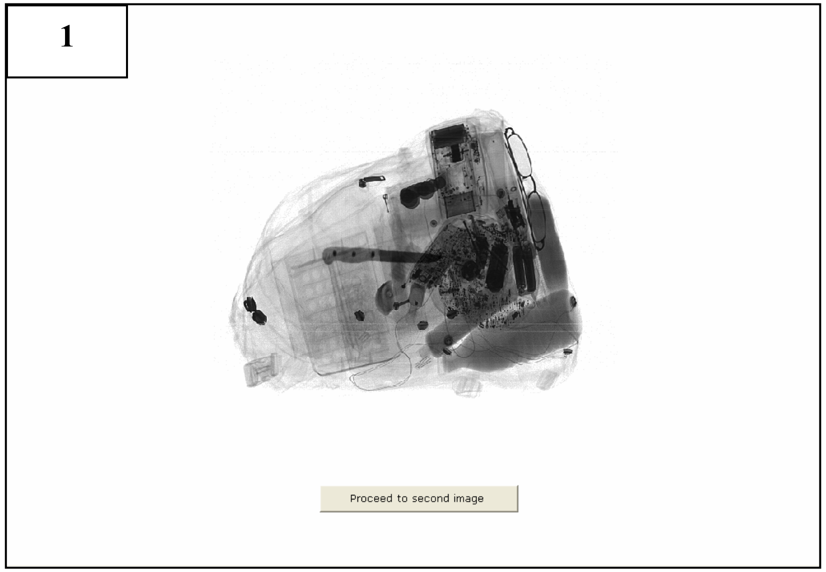
3.1.1 2AFC Experiment

Method

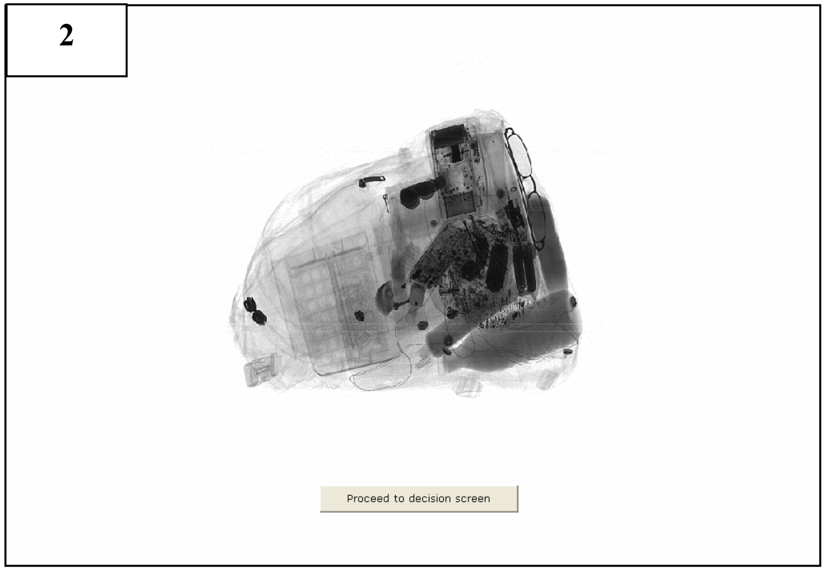
Data for the 2AFC pilot experiment were collected from 4 participants, ranging in age from 26 to 37 years old (M= 30.75, SD= 4.86). All participants were unpaid volunteers.

The presentation of the stimuli was sequential and participants were asked to choose if the first or second stimuli contained a threatening object. Participants were told that threats would be only knives. Thus it was up to the observer to discriminate which of the two competing stimuli contained the target. Additionally they had to rate the confidence of their decision on a 3-point Likert scale; either as ‘Certain’, ‘Probably’, or ‘Guessing’ (see Figure 10, boxes 1 to 5).

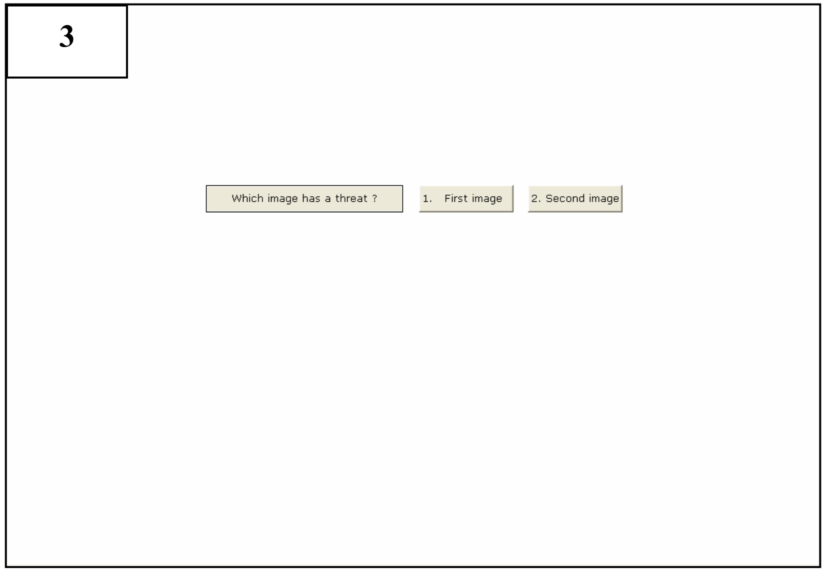
1


Proceed to second image

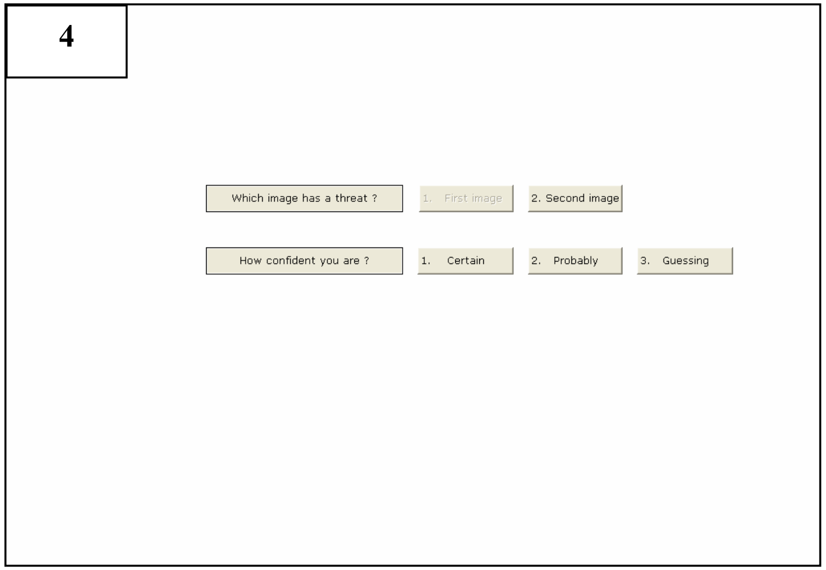
2

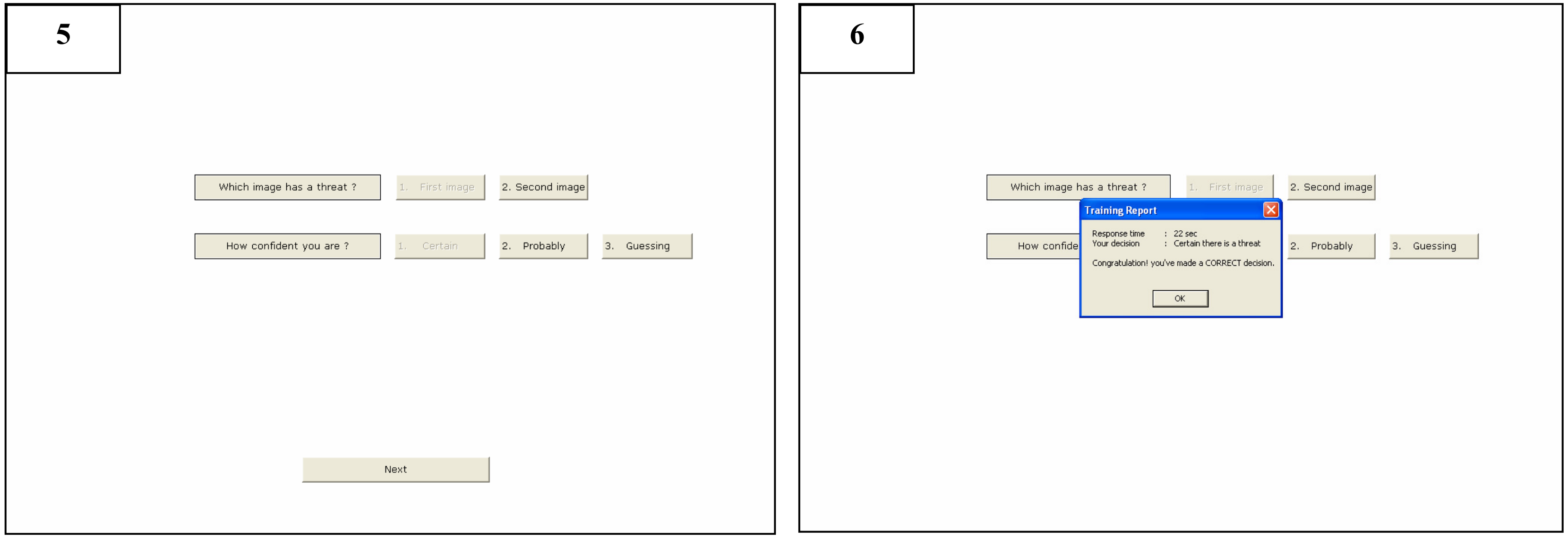

Proceed to decision screen

3


Which image has a threat ?
1. First image
2. Second image

4


Which image has a threat ?
1. First image
2. Second image
How confident you are ?
1. Certain
2. Probably
3. Guessing



Response time : 22 sec
Your decision : Certain there is a threat

Congratulation! you've made a CORRECT decision.

Figure 10. 2AFC pilot experimental procedure (boxes 1-5) and training feedback (box 6).

Note. Boxes: 1. first image presented; 2. second image presented; 3. is target in first or second image; 4. confidence rating; 5. go to next trial; 6. training feedback.

Prior to completing the experimental trials participants completed a short 10 trial training simulation intended to familiarize them with how the experiment worked and with the general visual properties of x-ray images of hand luggage. After each training trial the program provided the participants with instant feedback (see Figure 10, box 6), and their performance was discussed with the experimenter.

Overall 20 different pairs of bags were used, plus their mirror images; each presented in both viewing modes. Each image containing a threat was paired with its respective noise stimulus (i.e. no threat image), and was presented in 4 different presentation order and display mode combinations as detailed in the table below. Thus, the entire experiment was made up of 320 trials (pairs of images), with a 50% target prevalence. Threatening objects were small, medium or large knives (1 small, 1 medium, 1 large). The presentation of noise and signal trials was balanced so in half the trials the bags containing threats appeared first, and in the other half they appeared second. The presentation order of each stimulus pair was completely randomized, and then separated into 4 sessions of 80 trials each. Participants were allowed a break between each session, and completed the entire experiment on the same day.

Table 4. *Presentation order and stimulus conditions for the 2AFC pilot experiment.*

	Presentation Order			Presentation Order	
N	1st No Threat	2nd Yes Threat	N	1st Yes Threat	2nd No Threat
40	SM	SM	40	SM	SM
40	KM	KM	40	KM	KM
40	SM	KM	40	SM	KM
40	KM	SM	40	KM	SM

Note. SM: two-dimensional x-ray image, KM: kinetic depth effect x-ray image, Yes Threat: Image that contains a target (signal). No Threat: Image that does not contain a target (noise).

Results

Results from all participants were combined and analyses computed on the pooled data. Results were separately analysed for each of the four stimulus pairing conditions. Additionally, in the two cases where SM and KM stimuli are paired together two additional analyses were conducted to check if there was an interaction of display presentation order with temporal positioning of the target. Specifically, the data tables were arranged so the target was always in either the SM or the KM display.

As can be seen from the results (Table 5), when comparing between displays, participants were more sensitive in picking out the target in the KM display (d'= 1.34), as opposed to the SM display (d'= 0.75). Performance in the mixed display conditions was somewhere in the middle and was near identical (d'≈ 1) irrespective of which display the target was in.

Table 5. *Data analysis for the 2AFC experiment's different stimulus pairing conditions.*

Condition	H	F	d'	p(c)
SM-SM	0.59	0.21	0.75	0.69
KM-KM	0.78	0.13	1.34	0.83
SM-KM, SM-KM	0.72	0.18	1.07	0.77
KM-SM, KM-SM	0.69	0.18	0.99	0.75
SM-KM, KM-SM	0.72	0.18	1.05	0.77
KM-SM, SM-KM	0.69	0.18	1.01	0.76

Note. H = the hit rate, F = the false-alarm rate, d'= the index of sensitivity, and p(c) = the proportion correct.

Additionally, in the case of the SM-SM and KM-KM conditions a ROC analysis was conducted using the confidence ratings data. The results are presented below (see Table 6).

Table 6. *ROC data analysis for the 2AFC pilot experiment.*

Display Mode	H	F	d'	s	A_z
Static Mono (SM)	0.59	0.21	0.75	0.96	0.77
Kinetic Mono (KM)	0.78	0.13	1.34	0.48	0.90

Note. H= the hit rate, F= the false-alarm rate, d'= the index of sensitivity, s= the zROC slope, and A_z= the area under the ROC curve.

As expected the A_z results mirror the d' results. Specifically, the ROC analysis revealed an approximately 0.13 performance improvement in the KM display. This improvement is also evident in the following graph (see Figure 11). The SM line is closer to the chance line, whereas the KM line is more elevated indicating greater performance.

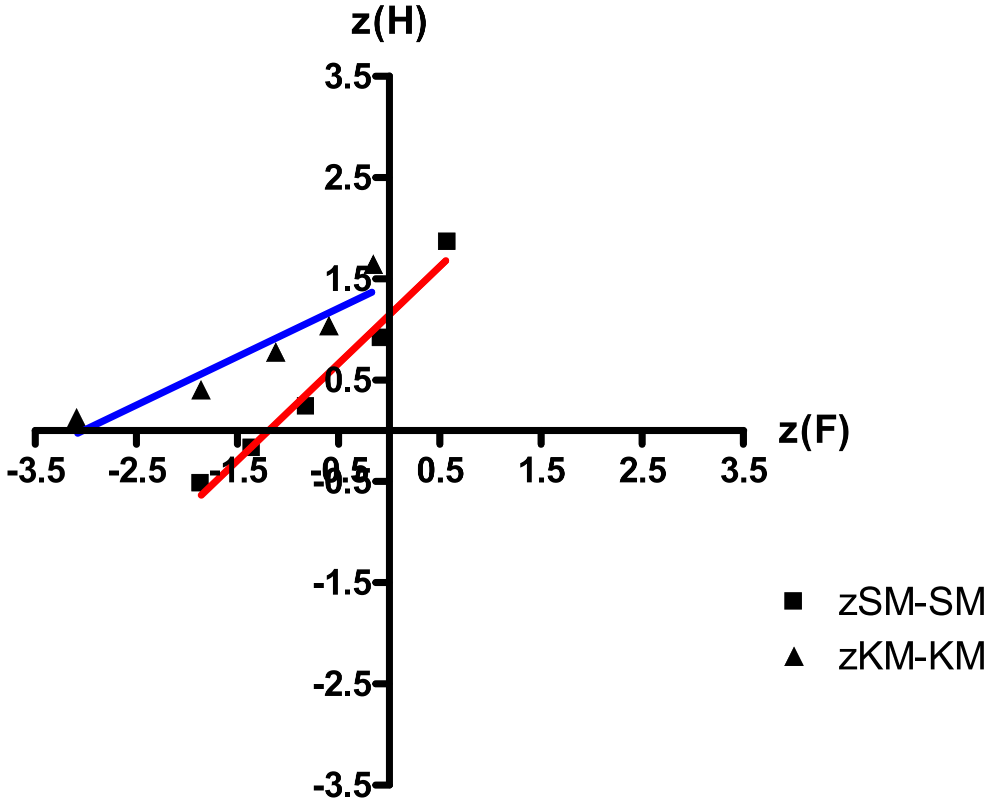


Figure 11. Pilot 2AFC experiment zROC for the SM-SM and KM-KM conditions.

Furthermore, the above graph reveals that the noise and signal distribution variances are nearly equal in the SM-SM condition, but unequal in the KM-KM condition (s<1). This is also confirmed by the respective slopes of 0.96 and 0.48.

Discussion

The 2AFC design was tested, but abandoned early in the course of the project. The main reason was the decision to focus on approximating, as close as possible in the lab, the actual process of threat detection as it happens in the field.

However there were valid reasons for initially considering this approach. One reason was because the 2AFC is ostensibly a more bias-free design than the Yes-No experiment (see SDT discussion in Chapter 2). This can guard against an observer bias to say “No” when dealing with a weak signal.

Another reason for experimenting with the 2AFC design was because it was hypothesized that it would be a more direct approach to ask participants to discriminate where the target is between two images; as opposed to trying to detect it in successive trials. Plus given the fact that the experiments were carried out using untrained volunteers it was felt that using a discrimination approach might make the task less daunting for them. Additionally, the 2AFC design is more appropriate in instances where the signal is very weak (see SDT discussion in Chapter 2).

In fact, this later point is reflected in the results. Participants in the 2AFC experiment performed higher in the SM display ($A_z = 0.77$) than the participants in the Yes/No experiment who viewed the same SM stimuli. This was also the case for the KM display ($A_z = 0.90$) as well. Furthermore, as can be seen participants were practically twice as good at discriminating between the signal and noise trials in the KM display (d'= 1.34), compared to SM (d'= 0.75).

However, although the 2AFC design assumes equal signal and noise distribution variances, the zROC slope (s = 0.48) for the KM-KM condition informs that they were in fact unequal. They were practically equal in the SM-SM condition (s = 0.96). This raises doubt about the validity of using the d' for describing this data.

In terms of the A_z there was a 0.13 improvement in performance for KM over SM. As was seen in the previous chapter, this was also evident by looking at the zROC data plots (see Figure 11).

Nonetheless, for all the above reasons, the 2AFC was not considered as the best fit for this model of airport security screening. Hence, it was discarded in favour of the Yes-No SDT design.

3.1.2 Yes-No Experiment

Method

Initially 11 participants were recruited for the study, but due to personal reasons, one of them did not complete the study. Consequently, his results are henceforth excluded. Complete datasets were collected from 10 participants (9 male, 1 female). All the participants were unpaid volunteers. They ranged in age from 25 to 37 years old (M= 28.80, SD= 3.94).

The Yes-No rating study utilized a test-retest counterbalanced design. The same 20 pairs of yes-threat no-threat images as the 2AFC pilot were used. Threatening objects were again different types of knives (1 small, 1 medium, 1 large). These 40 original images were presented in 4 different orientations, and the resulting 160 images were presented twice. Thus, the yes-no pilot study was also made up of 320 trials.

The presentation order of the images was randomised across the entire experiment, and furthermore the study was broken down in 4 sessions of 80 images each. This small pilot study has a near identical procedure to the next experiment (Experiment 1), which was in essence derived from it. Hence in order to avoid repetition a comprehensive description of the procedure will be withheld until the next section (see 3.2 Experiment 1: Threat Detection).

Results

The first step in the analysis was to separate the participants into four groups based on display mode (SM or KM) and experiment completion order (first or second). The results were then aggregated for participants in each subgroup and analyses conducted on the pooled data.

A ROC analysis was also conducted using the confidence rating data and the area under the linear ROC curve was calculated, as well as the slope of the standard score ROC (or zROC) for each subgroup. The results are summarised in the following tables (see Table 7 and Table 8).

Table 7. *Results from the first round of experiments*.

Experiment	H	F	c	s	A_z	log (β)	RT (sec)
Static Mono (SM) 1st	0.24	0.15	0.88	0.69	0.55	0.29	9.14
Kinetic Mono (KM) 1st	0.42	0.08	0.80	0.54	0.70	0.93	9.62

Note. H = the hit rate, FA = the false-alarm rate, c = the criterion, s = the zROC slope, A_z = the area under the ROC curve, log (β) = the natural logarithm of the likelihood ratio, and RT = the response time in seconds.

Table 8. *Results from the second round of experiments*.

Experiment	H	F	c	s	A_z	log (β)	RT (sec)
Static Mono (SM) 2nd	0.32	0.09	0.92	0.55	0.61	0.81	6.02
Kinetic Mono (KM) 2nd	0.49	0.13	0.59	0.54	0.68	0.66	8.67

Note. H = the hit rate, FA = the false-alarm rate, c = the criterion, s = the zROC slope, A_z = the area under the ROC curve, log (β) = the natural logarithm of the likelihood ratio, and RT = the response time in seconds.

The results indicate that in both experimental groups performance was better with the KM display. As can be seen from the data tables, in the SM-KM condition the KM display yielded a 0.13 improvement over SM (A_z of 0.55 and 0.68 respectively). Similarly in the KM-SM condition there was again a performance advantage of KM over SM, but of the order of 0.09 (A_z of 0.61 and 0.70 respectively). There are some hypotheses regarding why the performance improvement was lower for this group of participants, and these will be discussed later in the text.

To best visualize the ROC data obtained from the confidence ratings they were again plotted as standard scores (see Figure 12). These mirror the observations on the nature of the results made using the A_z scores.

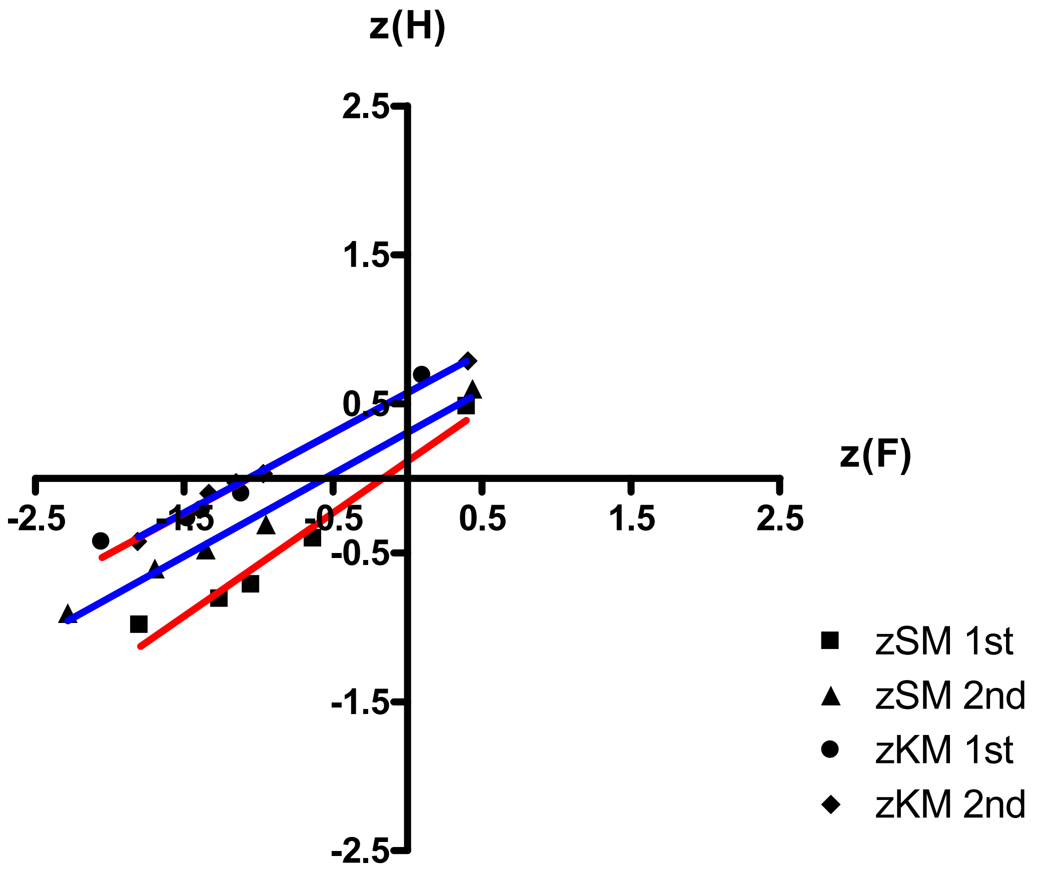


Figure 12. Pilot Yes-No experiment zROC across both experimental conditions.

By examining the zROC plot for the different groups it is evident that for the two KM conditions not only are the regression line slopes near identical, but the actual lines on the graph overlap. In the SM condition however there is a clear difference between the two groups. The line elevations reveal that performance for SM 2nd is clearly higher than SM 1st. In fact participants in SM 1st are operating close to chance as evident by the proximity of their data to the chance line.

Additionally, the slopes of the zROC lines can provide information with respect to the nature of the signal and noise distribution variances. As all the slopes are non unit (s<1) this indicates that the variances, of the signal and noise distributions, are unequal. This fact has implications on the selection of the appropriate performance index for the data. A lengthier discussion follows further on in the text.

Discussion

In the Yes-No experiment again a relatively strong advantage for KM over the SM display was found. With regards to the SM-KM condition, a 0.13 performance improvement was found as evident by the corresponding increase in the A_z values from 0.55 to 0.68.

The improvement was however reduced in the KM-SM condition with only a 0.09 improvement for KM ($A_z = 0.70$) over SM ($A_z = 0.61$). This is despite the fact that, in the case of the KM display, the data between the two groups of participants were virtually identical. In fact as can very clearly be seen in Figure 12 the two zROC lines, obtained via linear regression, overlap.

The suspicion is that a memory effect accounts for results obtained when participants viewed x-ray images in the KM display first, and the SM display second (KM-SM group). It is theorized that to a certain degree this might be due to the participants having had a better impression of the contents of the bags in the x-ray image in the KM display. Subsequently, when the same bag appeared in the SM display they either remembered having seen a target there before, or were sure that no target has been present. Both these explanations are possible as the hit rate was higher and the false alarm rate lower. This therefore might justify the higher than expected performance.

To satisfy this query a similar design was implemented in the next experiment (Experiment 1). As will shortly be seen, this enabled the estimation of the size of this effect, and to better ascertain its likely causes. By understanding this irregularity in the data steps can be taken to try and reduce it in the design of the future experiments.

3.1.3 Conclusions

These pilot experiments were exploratory in nature. They allowed the determination of most suitable SDT design for the visual search task entailing the inspection of x-ray images of hand luggage. It was concluded that the single trial design, as employed here, was best suitable for that purpose.

3.2 Experiment 1: Threat Detection

After conducting the pilot study, and selecting the single-interval (Yes-No) as the preferred SDT paradigm, the experiments were redesigned and the investigation proceeded with a more comprehensive study.

3.2.1 Method

Participants

Participants for the study were recruited from the University student and staff population, as well as the surrounding community. Only volunteers over 18 years old, with normal binocular visual acuity, were accepted. For this experiment 22 participants (10 male, 12 female) were recruited, ranging in age from 19 to 33 years old (M= 25.46, SD= 3.78). They were all given informed consent forms to read and sign before the experiment began, and their signed copies were retained. At the conclusion of the study they were compensated for their time with £40 in the form of gift vouchers.

Four of the participants had some previous experience with the procedures, having participated in an earlier pilot study. For the remainder this was their first encounter with the experimental tasks. However, for practical purposes they were all considered novices in terms of x-ray security screening (see Schwaninger, Hardmeier & Hoffer, 2005, for a discussion on the differences between expert security screeners and novices).

Design

The study utilizes a signal detection theory (SDT) framework. A variation of the basic yes-no SDT design, namely a two-response rating experiment (Macmillan & Creelman, p51-52, 2005) was employed. In this design, during each trial, participants are asked if a target is present or not, and then to rate the confidence of their decision (rating component). This design thus required two sub-responses from the participant during each trial (two-response component).

Participants completed trials in both display modes (SM and KM), but half of them viewed the static display first, and the other half the kinetic display first (see Table 9).

Table 9. *Experiment 1 participant groups, average age, and MARS scores.*

Participants	Age		MARS Score		Display Mode	
	M	SD	M	SD	Completed 1st	Completed 2nd
11 (5M, 6F)	24.82	3.49	1.82	0.04	Static Mono	Kinetic Mono
11 (5M, 6F)	26.09	4.11	1.81	0.05	Kinetic Mono	Static Mono

Note. F=Female, M=Male.

Following SDT procedure, when the target was present a positive response resulted in a hit (H), and a negative response in a miss. Similarly when the target was absent a positive response resulted in a false alarm (FA), and a negative response in a correct rejection. This enabled the computation of hit and false alarm rates for the dataset, which in turn made it possible to separate the participants' sensitivity from their bias. This is the main reason for using SDT as the mere comparison of the percentage of correct decisions does not permit assessing the effects of both bias and sensitivity on the results (Green & Swets, 1966).

Materials

The experimental sessions were carried out in the University group's laboratory, located at the Clifton Campus of Nottingham Trent University. The room dimensions were 2.6m x 4.9m, it had no windows and the walls were painted white. Two 20" (Iiyama HM 204 DT A) CRT monitors (set at a resolution of 1024x768 pixels with a 150 Hz Refresh Rate) were placed next to each other on a continuous bench. Each had a PC tower (Pentium 4, running Windows XP Home Edition) next to it, and a wireless keyboard and mouse was placed in front of each monitor. A 1.7m x 2.0m privacy screen was positioned between them, thus creating two identical experimental stations. As the following figure illustrates (see Figure 13), this setup enabled the simultaneous collection of data from up to two participants.

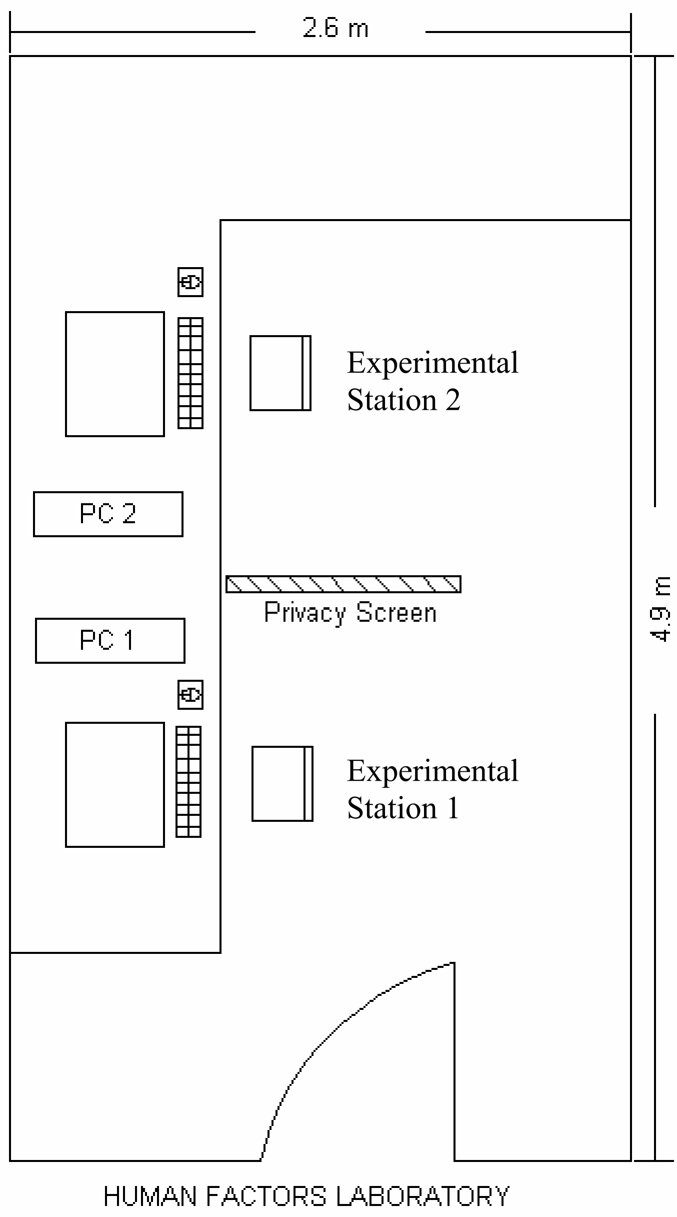


Figure 13. Layout of the laboratory.

Participants sat in front of the monitor, on an office chair. The height of the chair was adjusted so each participant's line of sight was approximately at the middle of the screen. The illumination inside the room was provided by reflective light from a single 40 watt incandescent bulb, and kept relatively constant throughout the course of the study.

Figure 14. Participant during testing session.

Experimental Stimuli

The images of different items of hand luggage were collected by the University group’s novel multiple view x-ray scanner (Evans & Hon, 2002; Hon & Evans, 2002). They varied with respect to background complexity, position of the target, and amount of overlap by other objects. Due to operational constraints (see beginning of Chapter 3) during image acquisition the x-ray images were presented in greyscale, in both the SM and KM conditions. Additionally, the threatening objects were limited to various types of knives (small, medium, large). Specifically there was 1 small, 1 medium, and 1 large knife.

Both the image presentation and data collection were entirely computerized. Bespoke software was specifically developed by the University group for this project (see Appendix A). The SM stimuli consisted of single frame x-ray images. The KM stimuli were kinetic x-ray images, composed of 16 different frames arranged into a continuous semi-rotation cycle; starting from the -7° frame, ascending to the +8° frame, and then descending back to -6° frame to complete the sequence (see Figure 15). Thus, each cycle of rotation was made up of 30 frames. Each frame was presented for 100 ms, with 1° of separation between adjacent frames, giving an angular speed of approximately 10 frames per second (or 10 degrees/second).

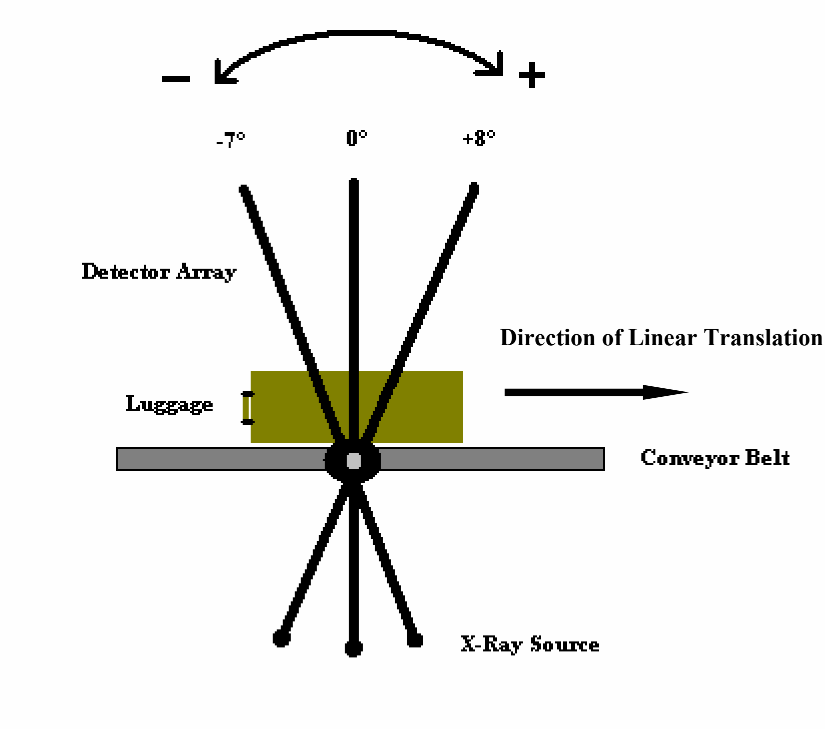
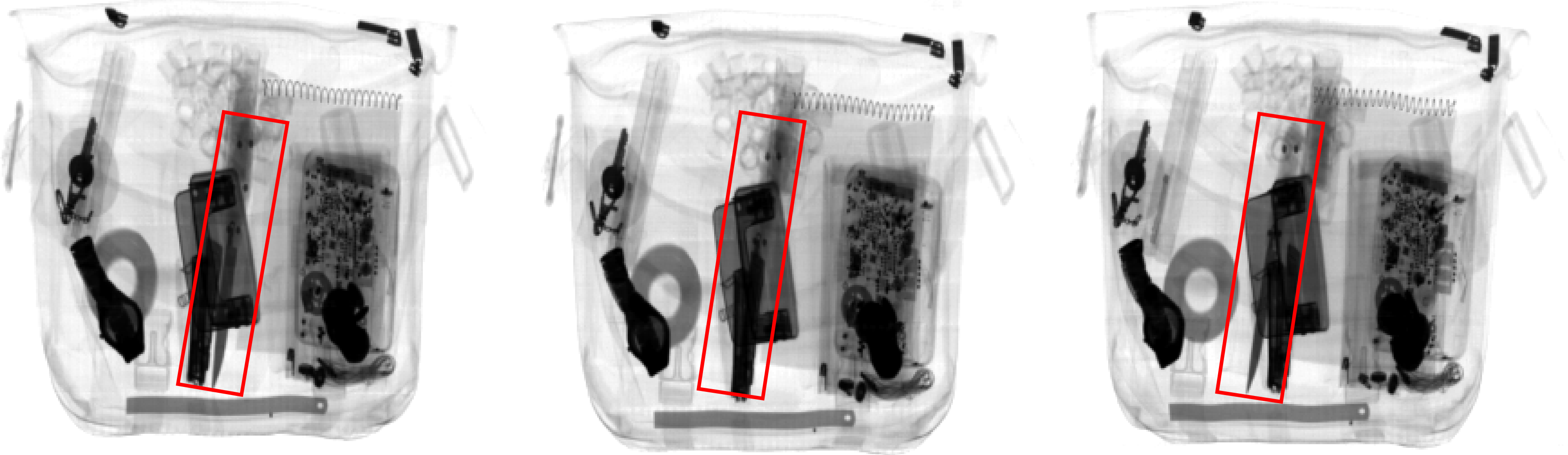


Figure 15. KM sequence explained.

A sample of the stimuli can be seen in Figure 16. In this scenario the target (knife) was placed inside a purse, crowded by other objects. The 0° image (frame 8) served as the SM stimulus. The other two images (frames 1 and 16) show either end of the KM sequence.

Overall the quality of the stimuli was good, however there were certain constraints associated with them. Specifically, as aforementioned, the effective angle of rotation of each bag was +/- 7 or 8 degrees. This is quite small as compared to the 180° or 360° rotation used by Wallach and O'Connell (1953) to demonstrate the KDE, but due to hardware limitations this is what could be achieved at the time. It is extremely important to note that this relatively small angle of rotation produces a noticeably weaker depth effect than that provided by wider angles of rotation (e.g. +/- 15° and above). Consequently, any advantage for the KM display, revealed in the present data, implies that even greater advantages may be encountered once the angle is increased.



FRAME 1 FRAME 8 FRAME 16

Figure 16. Sample x-ray images, with the target outlined.

Additionally, because the position of the target, relative to the plane of rotation, varied between stimuli this too constrained the angle of rotation of the target inside the bag. However, as stated earlier in the text, the ability of the human perceptual system to extract structure from motion is quite robust. Furthermore, there is evidence that observers can experience KDE even when confronted with an optical motion pattern that is heavily contaminated by visual noise (Todd, 1985). Hence, there is confidence in the present kinetic stimuli's ability to demonstrate the KDE. Nonetheless, a more thorough discussion on these constraints will follow later in the thesis.

It may also be worth noting that Wallach and O'Connell (1953) made use of the fact that the direction of depth in the third dimension is ambiguous in the KDE shadowgraph display. This manifests itself in the participant asserting that the direction of rotation abruptly switched from clockwise to anti-clockwise, and vice-versa. The presence of these reversals was taken as evidence that a genuine depth effect was produced in kinetic displays. In fact, where the edges or contours of shadows of objects did not change simultaneously in orientation and length, these reversals were not observed. In point of fact, despite the relatively small angle of rotation used, the apparatus also yielded easily observed reversals of the direction of rotation.

Two classes of stimuli were employed in the experiment: images that contained a threat (Yes-Threat Image), and images that contained no threat (No-Threat Image). The 10 different bags used, were presented in four different spatial orientations. Each of the 40 resulting yes-threat images (containing a knife) was complemented by a no-threat image. These consisted of the same bags as used in the other set, but with the knives removed. The resulting 80 images were presented 4 times during the course of each experimental condition. Thus, an entire experimental block was made up of 320 trials for each display mode; where half the images contained a threat, and the other half did not. All of these images were presented in an intermixed random order.

Participants completed each block in 4 sessions of 80 trials. Each session contained one full set of images. Since participants completed one such block for each display mode they each performed a total of 640 trials during the entire study.

Visual Testing

At the outset, two short visual tests were administered (see Appendix B). The first was a visual acuity test in the form of a mini Snellen chart mounted on the wall, and observed from a distance of 1.83m (6 feet). The Snellen test is the undisputed standard in visual acuity testing.

The second was the MARS letter contrast sensitivity test (Arditi, 2005). This test was performed to ensure that participants had normal contrast sensitivity; an important prerequisite given the nature of the stimuli (greyscale images). The Mars test has been independently shown to have good reliability and validity (Dougherty, Flom, & Bullimore, 2005; Haymes et al., 2006; Thayaparan, Crossland, & Rubin, 2007).

The MARS test included three boards with different letter combinations (see Appendix B). The letters reduce in contrast sensitivity, and the participants were asked to go through each successive line and report what letter they were seeing. The testing continued until they made two consecutive errors or they had reached the end of the board. They were tested at a distance of 50cm from the board using their right eye, left eye and binocularly. As per the instructions, the correct illumination of the board had to be at least 85 cd/m^2 and uniform throughout the surface of the board. Both conditions were met by using an additional light source (60 watt incandescent light bulb), placed at approximately 45° angle to the board, and verified by a photometer.

Training

Each experiment day was preceded by a training session, composed of 10 practice trials, each providing the participant with feedback. Half of the practice trials contained a threat (in this case a knife) while others did not. This was intended to familiarize them with airport scanner x-ray images, and more importantly how threatening objects, as well as commonplace travel items, appear in these types of images. Specifically, it was explained to participants that the way the various items appear in an x-ray image depends on their material composition. For example, denser objects (e.g. metals) appear darker, and less dense objects (e.g. wood and soft plastics) appear lighter. Clothes are typically almost transparent, and when an object is black (typically the densest of metals) then it is opaque to x-ray beams.

The practice stimuli were different from those presented during the experimental trials. However the procedure was identical (with the notable exception of the feedback provided). A detailed description of the experimental procedure follows in the next section. An example of the stimuli used for training can be seen in Figure 17 (box 5), as well as an example of the feedback screen (box 6).

Once the participants had acquired the task, they took a brief break and then proceeded with the regular experimental trials, without feedback, and their responses were recorded.

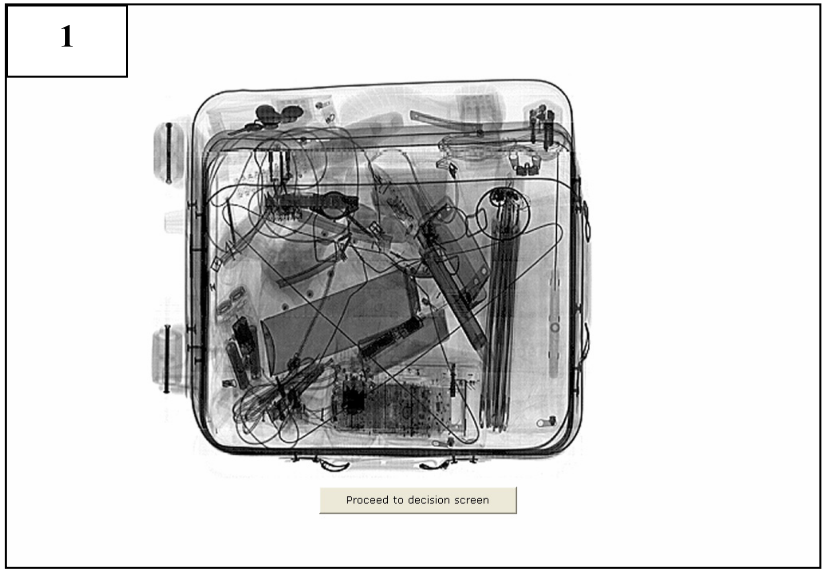
Experimental Procedure

As aforementioned, participants were divided into two groups, based on the completion order of the two display modes. It was also ensured that the four participants with some prior experience were equally divided between the two groups. The study took place over two days, with a break of about a week between the first and second day. Each of the experimental sessions lasted between 20-30 minutes, with a short break in-between. This duration was in line with what other researchers have reported using (e.g. Washburn, Taglialatela, Rice, & Smith, 2004).

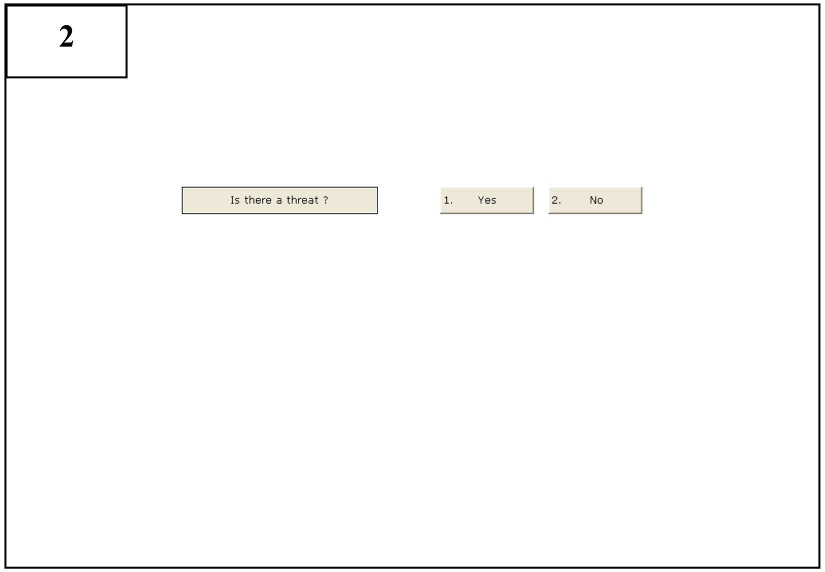
Participants were handed the keyboard, and shown the relevant keys: spacebar and numerical keys 1, 2 and 3. As a general rule participants were instructed to react as quickly and as accurately as possible, and their response time during each trial was recorded. The participant was instructed to pay attention to each x-ray image (see Figure 17) that appeared on their screen, and when ready press the spacebar in order to proceed to the decision screen. It was explained to the participants that potential targets would only be knives.

Once at the decision screen they were asked to decide if the preceding image contained a threat or not, and to qualify their decision by pressing “1” for “Yes” or “2” for “No”. Afterwards they were prompted to give a confidence rating on a 3-point Likert scale (1=”Certain”, 2=”Probably”, 3=”Guessing”). Following that, participants were required to press the spacebar in order to proceed to the next trial, at which point they were instructed to proceed through this and each subsequent trial as before.

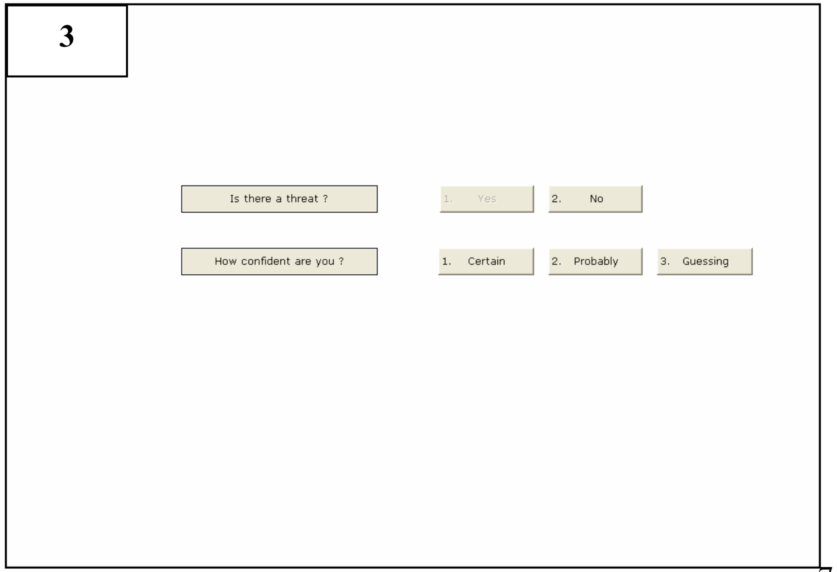
1


Proceed to decision screen

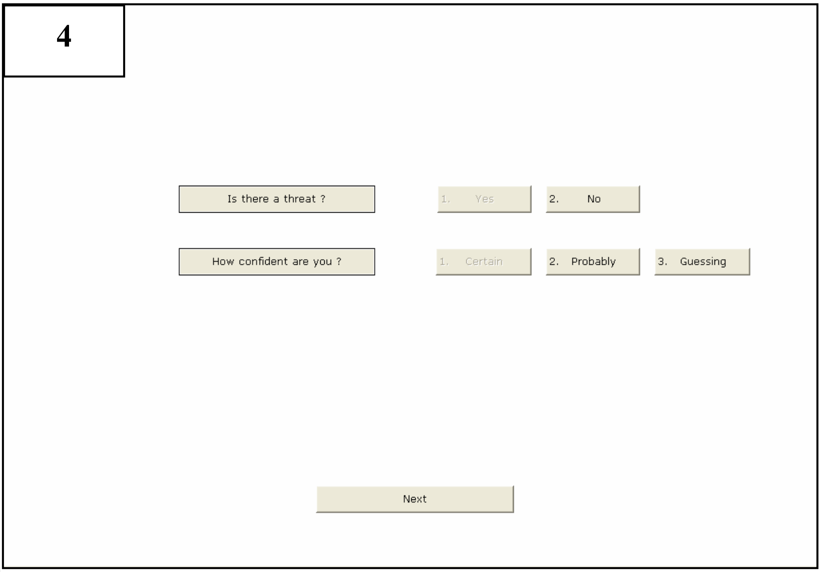
2

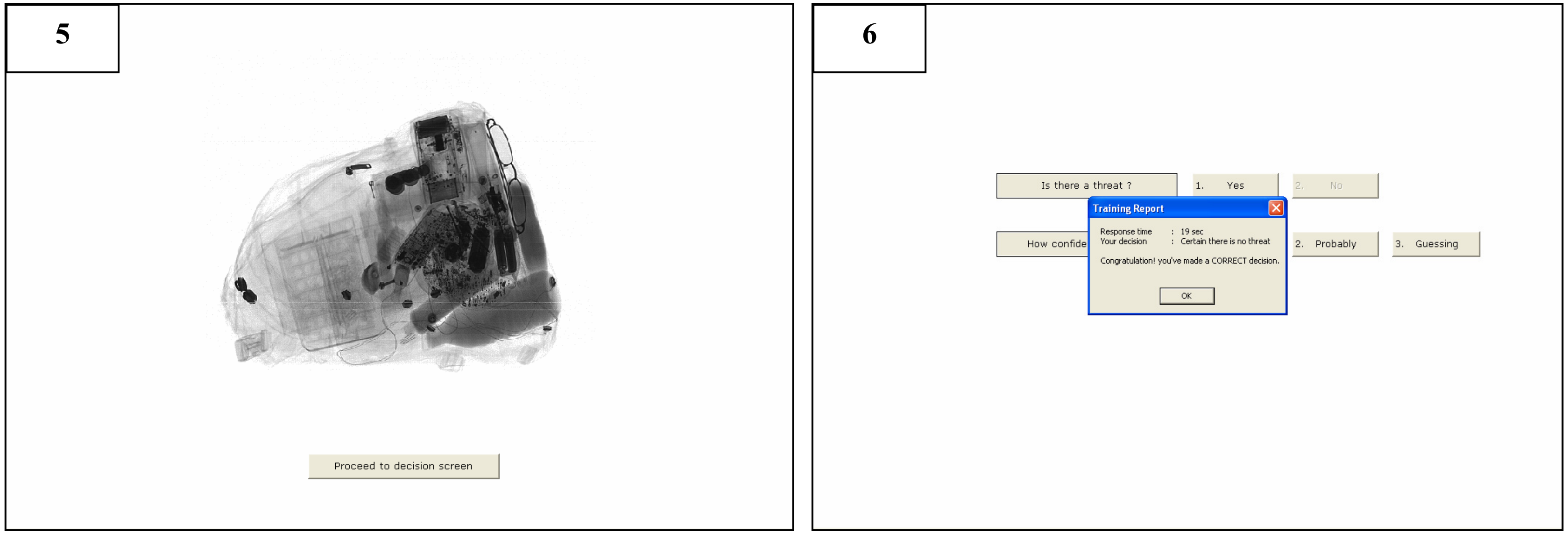

Is there a threat ?
1. Yes
2. No

3


Is there a threat ?
1. Yes
2. No
How confident are you ?
1. Certain
2. Probably
3. Guessing

4


Is there a threat ?
1. Yes
2. No
How confident are you ?
1. Certain
2. Probably
3. Guessing
Next



Training Report

Response time : 19 sec
Your decision : Certain there is no threat

Congratulation! you've made a CORRECT decision.

Figure 17. Experiment 1 (Threat Detection) procedure (boxes 1-4) and training (boxes 5-6).

Note. Boxes: 1. experiment image presented; 2. yes or no threat decision; 3. confidence rating; 4. go to next trial; 5. sample training image; 6. training feedback.

The instructions to the participants were identical for both experiment days. At the conclusion of the study the participants were debriefed, thanked for their involvement, and given their compensation.

3.2.2 Results

Visual Tests

The Snellen acuity test revealed that all of the participants had at least normal or corrected binocular acuity of 20/25. The results on the MARS test were entirely within normal parameters for all the participants and furthermore the two groups were similar in their average log contrast sensitivity. Their respective binocular scores were M = 1.82, SD = 0.04 for the SM-KM group, and M = 1.81, SD = 0.05 for the KM-SM group. The difference between the two groups was not statistically significant, t(20) = 0.55, p = 0.59.

Response Time

The average response times (RT), in seconds, were calculated for each display mode in the two experimental groups. The results for the SM-KM group were M = 12.61, SD = 3.39, and M = 9.40, SD = 2.94 for SM 1st and KM 2nd respectively. In the KM-SM group obtained figures were M = 13.83, SD = 3.74 for KM 1st, and M = 9.73, SD = 3.44 for SM 2nd.

Additionally a repeated measures analysis of variance (ANOVA) was performed (with display mode as the within group factor, and completion order as the between group factor), comparing response times across the first and second part of the experiment. No significant effects were found for either display mode or completion order (F < 1). However there was a significant interaction between the two F(1, 20) = 29.63, p < 0.001. This indicated that participants, in both SM-KM and KM-SM experimental groups, had significantly lower response times in the second part of the experiment irrespective of display mode. This is probably due to the repetition of the stimuli and further discussion will follow.

Furthermore two t-tests were performed to investigate the differences in RT between the two experimental groups (SM-KM and KM-SM), but within the same display type (SM and KM). As these tests were post hoc, in order to reduce the possibility of a false positive significance result (Type I error) the p level for significance was set at a more conservative p = 0.01, instead of the customary p = 0.05. The difference was found to be significant for KM, t(20) = -3.09, p < 0.001, but not for SM, t(20) = 1.98, p = 0.62.

ROC Analysis

As this was the first major experiment it was decided to also perform significance tests on the results. Hence, separate signal detection theory analyses were conducted on the data from each participant, and then averaged across display mode and experimental group. Based on this initial analysis, it appeared participants in the SM-KM group were exhibiting close to chance performance in the SM 1st condition with a relatively low hit rate (M = 0.59, SD = 0.13), and a relatively high false alarm rate (M = 0.42, SD = 0.22). Thus they showed increased uncertainty. In contrast in the KM 2nd condition the hit rate climbed (M = 0.73, SD = 0.16) and the false alarm rate declined (M = 0.16, SD = 0.15). Results for the KM-SM group show a similar trend for KM 1st with a comparable hit (M = 0.70, SD = 0.12) and false alarm rate (M = 0.16, SD = 0.07). However, for SM 2nd although the hit rate was close to chance (M = 0.53, SD = 0.10), the false alarm rate was relatively low (M = 0.16, SD = 0.12). Thus, in that instance uncertainty appeared reduced.

In order to conduct more robust comparisons, a receiver operating characteristic (ROC) analysis was conducted using the confidence ratings data. By plotting the hit rate (vertical axis) against the corresponding false alarm rate (horizontal axis), each participant's performance could be tracked across the decision spectrum (cf. Macmillan & Creelman, 2005, p51-57; Wickens, 2002, p84-92, for a detailed description of the technique for obtaining 6 data points from 3 levels of confidence).

Because for practically all participants the slopes of the z-coordinates ROC were non-unit (s < 1), the area under the curve (A_z) was deemed to be the appropriate performance index for the present data (Macmillan & Creelman, 2005, p63-64). The A_z is generally accepted as providing a good index of sensitivity when dealing with this type of data. It is determined by measuring the area under the linear coordinates ROC curve. This measure varies in value between 0.5 and 1.0; which respectively correspond to a total lack of sensitivity (chance performance), and maximum sensitivity.

For the SM-KM group A_z values of 0.61 (SM 1^{st}) and 0.82 (KM 2^{nd}) were obtained. Thus a 0.21 improvement for these participants was observed. In the KM-SM group the respective values were 0.81 (KM 1^{st}) and 0.71 (SM 2^{nd}), so the improvement was reduced to 0.10. Possible explanations for this difference are discussed later on in the thesis. The results of the data analyses performed are summarized in the following tables.

Table 10. *Data analysis for the SM-KM experimental condition.*

Display Mode	H		FA		A_z		RT (sec)	
	M	SD	M	SD	M	SD	M	SD
Static Mono (SM) 1^{st}	0.59	0.13	0.42	0.22	0.61	0.06	12.61	3.39
Kinetic Mono (KM) 2^{nd}	0.73	0.16	0.16	0.15	0.82	0.09	9.40	2.94

Note. H = the hit rate, FA = the false-alarm rate, A_z = the area under the ROC curve, and RT = the response time in seconds.

Table 11. *Data analysis for the KM-SM experimental condition.*

Display Mode	H		FA		A_z		RT (sec)	
	M	SD	M	SD	M	SD	M	SD
Kinetic Mono (KM) 1^{st}	0.70	0.12	0.16	0.07	0.81	0.07	13.83	3.74
Static Mono (SM) 2^{nd}	0.53	0.10	0.16	0.12	0.71	0.05	9.73	3.44

Note. H = the hit rate, FA = the false-alarm rate, A_z = the area under the ROC curve, and RT = the response time in seconds.

The A_z data were subjected to a repeated measures ANOVA, with display mode as the within group factor, and completion order as the between group factor. Results indicated a significant effect for display mode $F(1, 20) = 170.56$, $p < 0.001$. Hence, the performance difference was significant for both the SM-KM, and KM-SM conditions. Furthermore, there was a significant interaction between display mode and experimental order $F(1, 20) = 23.29$, $p < 0.001$. However, as expected, there was no significant effect for completion order $F(1, 20) = 3.91$, $p = 0.062$.

As the effect of completion order was not by itself significant, this variable could not by itself account for the difference in performance gain between the two experimental groups. The significant interaction between display mode and completion order however indicates that the order in which trials, in the two different display modes, were completed had at least a partial effect on performance.

Performance on each display type, across the experimental groups, was also examined. Again, as these tests were post hoc, in order to reduce the possibility of a Type I error, p level was set at 0.01. Differences were found to be significant for the SM display, $t(20) = -4.81$, $p < 0.001$, but not for the KM display, $t(20) = 0.20$, $p = 0.85$. This latter finding further hints that a repetition effect may have resulted from the completion order. This suspected repetition effect did however seem to be limited to the KM-SM group and hence possibly skewed the SM2nd data. This observation will be further discussed later in the text.

Pooled Sensitivity

Additionally, in order to visually display and examine the ROC data results in each display mode were combined across participants, separately for each experimental group, and analyses conducted on the pooled data. To analyze and graphically present the results Prism 4 (Graph Pad Software, 2004) was used. The pooled ROC data were plotted in z-coordinates and linear regression was employed to obtain the slopes (see Figure 18). In the SM-KM group the slope for SM 1st was $s = 0.76$, $F(1, 3) = 295$, $p < 0.01$, and for KM 2nd $s = 0.36$, $F(1, 3) = 27.16$, $p < 0.05$. In the KM-SM group, for KM 1st a slope was obtained of $s = 0.37$, $F(1, 3) = 36.78$, $p < 0.01$, and for SM 2nd $s = 0.39$, $F(1, 3) = 101.70$, $p < 0.01$. All resulted in a good linear fit; thus confirming the assumption that the underlying distributions are normal (Gaussian).

The elevations of the slopes of the zROC curves in Figure 18 enabled the visual comparison of performance across display mode and experimental group. This comparison reveals that performance was highest in the KM 2nd followed by the KM 1st condition. However this difference is marginal, as the two lines practically overlap. The graph also reveals that participants performed higher in the SM 2nd condition, as compared to the SM 1st, the performance of which was much closer to the major diagonal (chance line). These observations mirror the results based on A_z discussed earlier.

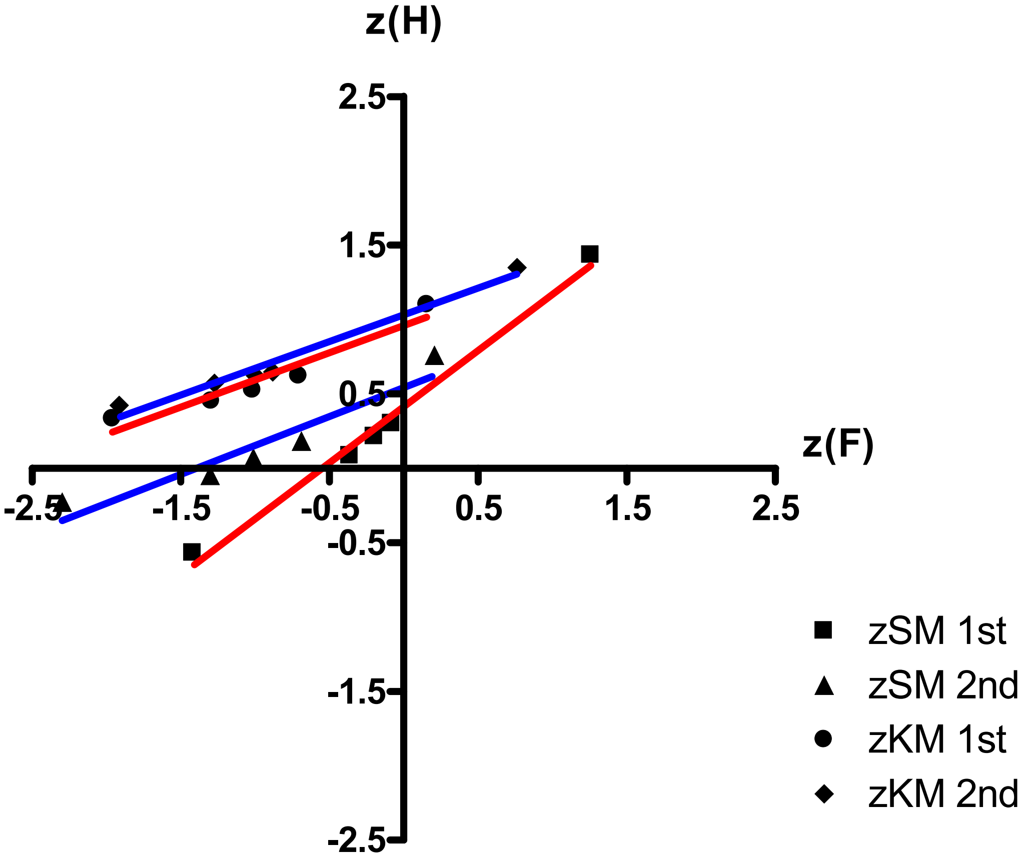


Figure 18. Standardized scores ROC of SM and KM display data across both experimental conditions.

Furthermore, the signal (plus noise) distribution variance was found to be greater than the noise distribution variance. This is consistent with the non-unit nature of the slopes in the zROC graphs.

3.2.3 Discussion

It is evident from the experimental results that despite the limited range of angular motion, KM x-ray images offer a superior rate of detection, as compared to SM x-ray images. Overall, both groups of participants showed a significant performance enhancement for kinetic, as compared to static displays. However, there were differences in the performance gains between groups. This was not due to a decrease in their performance in the kinetic condition, but rather due to a higher than anticipated performance in the static condition. In fact by revisiting the results it can be seen that performance in the kinetic condition was very similar for both groups of participants.

This leads to the conclusion that any effect of repetition on performance was limited to the SM condition of the KM-SM experimental group.

It is noteworthy that Wallach, O'Connell and Neisser (1953) reported a memory effect of the KDE in which participants viewed statically presented images after viewing shadows of rotating wire frame images. A consequence of prior viewing of KDE- producing images was that the participants tended to describe the static images as three dimensional. While it is tempting to interpret the asymmetrical results of the experiment where targets are more readily detected in static images after viewing kinetic images, it may well be that having seen images in depth created a general bias to interpret any similar but static image as three dimensional, thereby seeking out and attending to cues to depth (e.g. partial occlusion in the static displays). Before reaching any conclusions with regard to the memory effect of the KDE, and its relevance to the present study, more experimental work should be conducted.

An alternative explanation for this apparent improvement in SM performance of participants in the KM-SM condition is as follows. It may well be that the asymmetry can be best attributed to spatial implicit learning, where the participants having "seen" the revealed threat in the KM display were able to improve their responses. By revisiting Figure 16, it can be observed that while the target was crowded by other objects in SM, thus possibly evading detection, it was much more salient in KM. This explanation is also consistent with comments made by participants during their debriefing (such as remembering having seen a target in KM being revealed behind a masking object).

Furthermore, as was seen earlier, this same effect was also encountered in the initial pilot study conducted prior to this one. As already stated, this issue may only be clarified in the context of additional purposely designed experiments. The priority however, is to modify the future experimental designs so as to limit the possibility of any repetition effects, and focus instead on the main task of further evaluating the potential of the KM display.

In addition, it might be argued that the relative superiority of the kinetic mode is due to the simple fact that participants get to view any particular image from several different station points. Thus, in effect participants are exposed to many different versions of the same stimuli and can use the resulting information to mentally construct a representation of the three-dimensional objects contained in a given bag. Thus, the superiority may not be due to the kinetic depth per se, but to the aforementioned increased opportunities to gather information about the contents of a given bag. However, to the author's knowledge there is no evidence in the literature to support this view.

Finally, the kinetic depth effect depends upon the succession of changing lengths and directions of the contours describing the objects. Motion parallax per se does not necessarily incorporate these features. It is possible to have KM images, where the target and the mask (e.g. knife and umbrella) are both either horizontal or vertical (their main axes correspondingly parallel or perpendicular to the direction of conveyor belt motion), but at different depths. The proviso here is that the knife and umbrella are both considered approximately planar in shape (i.e. their depth is small relative to the depth discrimination of the system/observer). Thus, in the above scenario there is sequential parallax, but no KDE, as the latter also requires changes in contour orientation and length. This makes the timing and order in which the distribution of stimulus elements (features) are presented, in the KM display, of paramount importance for generating the KDE.

However the counter argument can very well be made that the patent depth perception, made possible by the KM display, facilitates the segregation of a target from its background and leads to the perception of a more readily recognized object. Thus, enabling participants to pick out threatening objects that would be otherwise lost in the clutter.

In future empirical work all these conjectures will be explored using much improved imagery. Since a significant performance improvement is currently being picked up with KM, using only a small angle of rotation, it is plausible to suggest that a + / - 20° angle would give a much greater improvement. Additionally, the plan is to use a greater number, and variety of images. Subsequently, the results obtained using vertical images and horizontal images (parallax but no KDE) may be segregated from oblique images (which give KDE). This qualitative analysis may definitively inform if the KDE is responsible for all or most of the variance in improved detection ability, or if parallax per se contributes to this variance.

Additionally, the intention is to migrate from greyscale to pseudo-colour images, using the industry standard colour coding scheme based on atomic number (cf. Sobania, & Evans, 2005). This opens up the possibility of determining if the KDE offers any special, but as yet unascertained, advantages based on changing colours of partially overlapping objects of different relative atomic numbers. The counter argument is that as different materials will appear in different colours there can potentially be target segmentation in objects that are composed of diverse materials. For example, in the case of a knife, the blade can be blue or black, and the handle orange or green. Hence pseudo-colour may work both ways; either productively by improving integration of "features" as in shifting of colour mixtures with motion, or counter-productively by promoting false impressions of several objects. This was the primary impetus for one of the later experiments which attempted to answer these very questions.

With respect to response time, there were no overall differences between the two groups. Display mode did not appear to affect response time; in fact the mean response times between the two experimental groups were nearly identical. A general trend of shortening of RT during the second day of testing for both experimental groups was however observed. This pattern of shortening of response times, with no discernible change in sensitivity, was an anticipated consequence of the research design and is adequately explained by the contextual cueing effect (Chun & Jiang, 1998), a form of implicit learning in visual search tasks. Thus, it can confidently be stated that this overall reduction stemmed from the experience gained in the first round of experiments.

In summary, the results obtained in this experiment demonstrate that the KDEX imaging technique shows great promise in addressing the problems associated with static images. The main benefit is overcoming the inability of current imagery to show depth, thus providing a better and cost effective imaging alternative. Other associated limiting factors such as ordinary hiding of targets by occluding objects, and crowding, where background objects distort the target and prevent feature integration (Pelli, Palomares & Majaj, 2004), may well be partially overcome by this new technique.

This being the first major experiment, it helped provide answers to several questions formulated during the pilot study. However as is the nature of research, it also ended up generating a new set of unanswered questions. During the discussion of the results several questions were also posed. These influenced the design and objectives of subsequent experiments. For instance in the next experiment it was decided to alter the design and also incorporate an identification component in the procedure.

3.3 Experiment 2: Threat Identification

In the previous experiments it was demonstrated that KM x-ray images afford a better chance for threat detection, compared to SM images. This experiment went one step further and set out to establish that KM images also lead to the better identification of threatening objects. The approach taken was similar to the LROC technique described earlier in Chapter 2.

In deciding on the design of the identification component other simpler paradigms were considered. For example, a classification experiment where the observer would be required to first indicate if a target is present or not and then identify which category it belongs to out of a predetermined list; for instance knife, gun, scissors or explosives. This approach has practical similarities in its application to source monitoring experiments (cf. DeCarlo, 2003) where an observer has to first decide whether a target is present and then discriminate the source of the target. Obviously the observer would only proceed to the second stage if he or she detected a target in the first place.

Nonetheless since it was technically feasible to have participants directly selecting what they perceived as a target, and for their responses to be automatically recorded, it was decided to adopt this more complicated, but also more promising approach.

3.3.1 Method

Participants

Participants for the study were again recruited from the student and staff population of NTU, and the surrounding community. In addition to printed advertisements posted in various locations the experimenter also created a project website. For this experiment 18 participants (15 male, 3 female) were recruited, with normal (or corrected) 20/25 visual acuity. They ranged in age from 18 to 38 years old (M= 24.33, SD= 4.83), and were compensated for their time with £20 worth of gift vouchers. All were first time participants, with the exception of one 34 year old male who also took part in the pilot study.

Table 12. *Experiment 2 participant groups, average age, and MARS scores*.

Participants	Age		MARS Score		Display Mode	
	M	SD	M	SD	Completed 1st	Completed 2nd
9 (7M, 2F)	24.22	5.74	1.79	0.05	Static Mono (A)	Kinetic Mono (B)
9 (8M, 1F)	24.44	4.07	1.82	0.06	Kinetic Mono (A)	Static Mono (B)

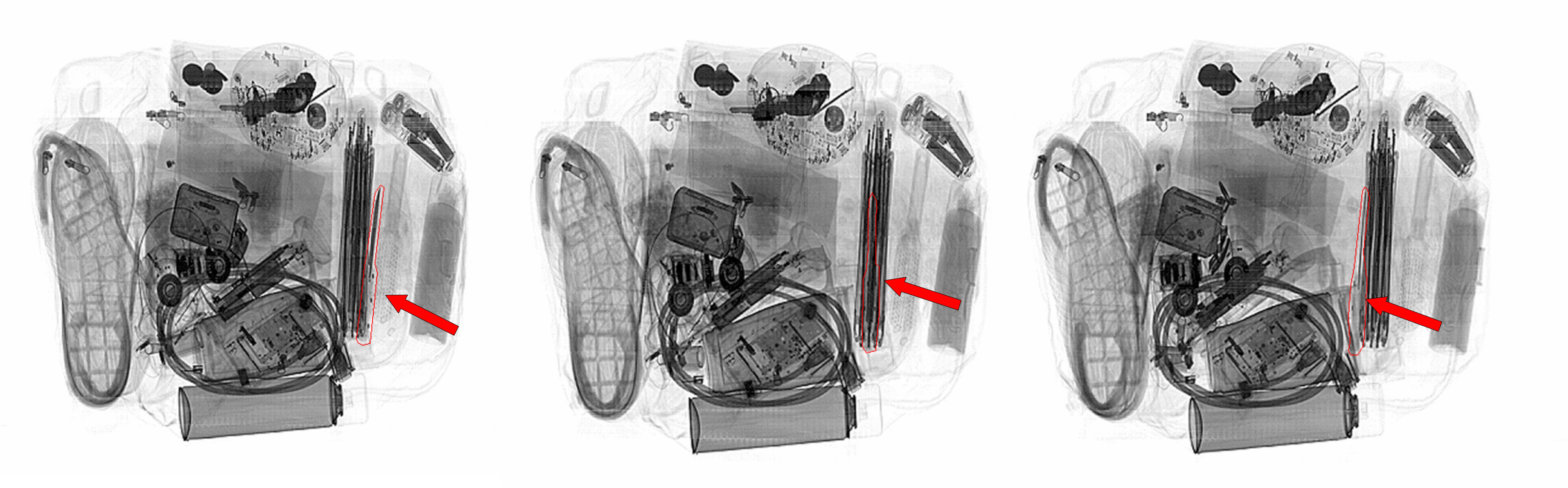
Note. F=Female, M=Male.

Equipment

The same laboratory facilities as in the previous experiment were used. As before, the experiments were entirely computer based, and stimulus presentation, and data collection was accomplished using a modified version of the same custom software, which was previously developed by the University group for this project. Again, participants sat approximately 45-50cm away from a 20” CRT monitor (resolution 1024x768 pixels, refresh rate 150 Hz), with their line of sight at about the middle of the screen. The room was illuminated with a single 100 watt incandescent light bulb.

Experimental Stimuli

Greyscale x-ray images with varying degrees of background clutter, and target visibility were used. As with the previous experiments threatening items were present in 50% of the images, and the kinetic stimuli were composed of 16 frames. Additionally, as with the preceding experiments targets were exclusively different types of knives. Specifically, there were 1 small, 1 medium and 1 large different knives. In total images of 19 unique bags were used; one bag appeared in both experiments. Each threat-present bag image was complemented by its respective non-threat image, and presented in 4 different orientations. The bags were presented 4 times, during the course of the experiment, resulting in a total of 320 trials. Participants completed the experiment in 4 sessions; each consisting of 80 randomly presented images.



FRAME 1

FRAME 8

FRAME 16

Figure 19. Sample x-ray images, with the target area outlined in red.

Note. A click within this area was considered a hit in the identification part of the experiment.

Each target area was traced in MS Paint and threat location spatial maps were generated listing the x and y screen coordinates containing each threatening item. This enabled the computer program to automatically log if an object selected by the observer, as a threat, was genuinely a threat or not (see Figure 19).

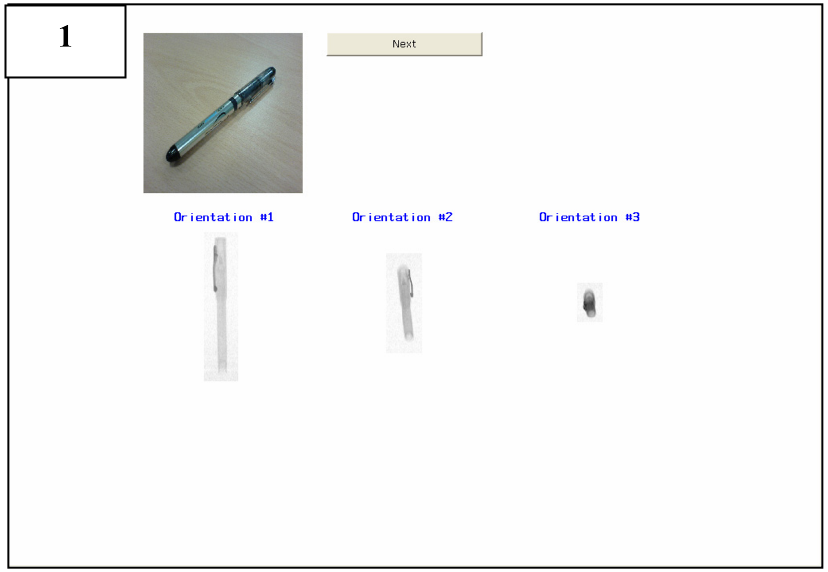
Training

Participants received a two-part training course. In the past the only training given was the experiment practice simulation with feedback before the commencement of actual experimental trials. While this was crucial to the participants acquiring the experimental task the experimenter felt that they needed further exposure to x-ray images before undertaking the screening experiments. Hence it was decided to look for ways to improve participant training as much as time permitted.

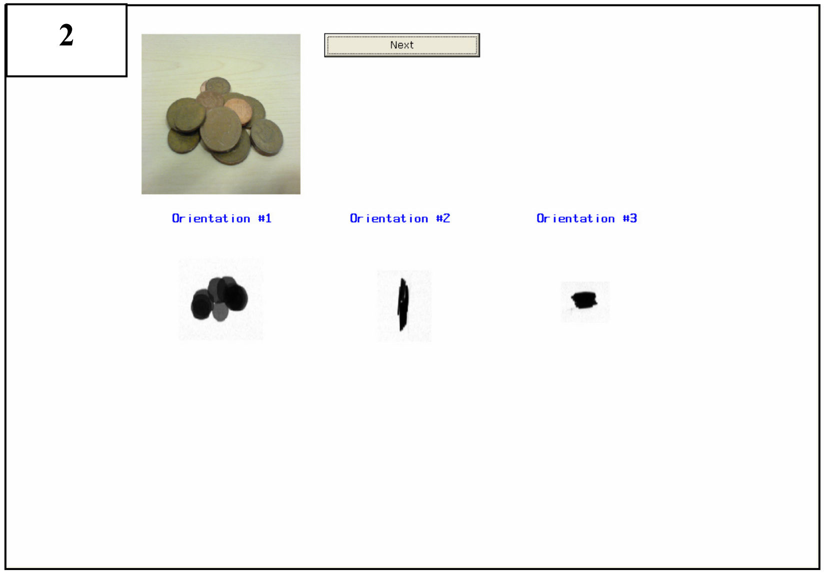
Therefore the participants in this experiment first received a new round of stimulus familiarization training designed to first introduce them to how objects appear in x-ray images. This was accomplished by designing a training program that showed them around 20 pictures of objects in visible light as well as three x-ray orientations. The objects were those of typical travel items. They increased in complexity, in terms of the materials they were made of, during the course of the training program. The objects displayed included 3 threatening items (scissors, knife and pistol).

Subsequently, participants were shown 3 images of a synthesized bag filled with an assortment of some of the 20 objects they previously viewed. Each bag also included one of the previously shown threatening objects. Furthermore, there was an increasing degree of overlap between the different objects. Participants were encouraged to identify each object and click on it using the mouse cursor, which then displayed the name and picture of the corresponding object on the program screen. Figure 20 contains some examples of the images participants viewed as part of their stimulus familiarization training.

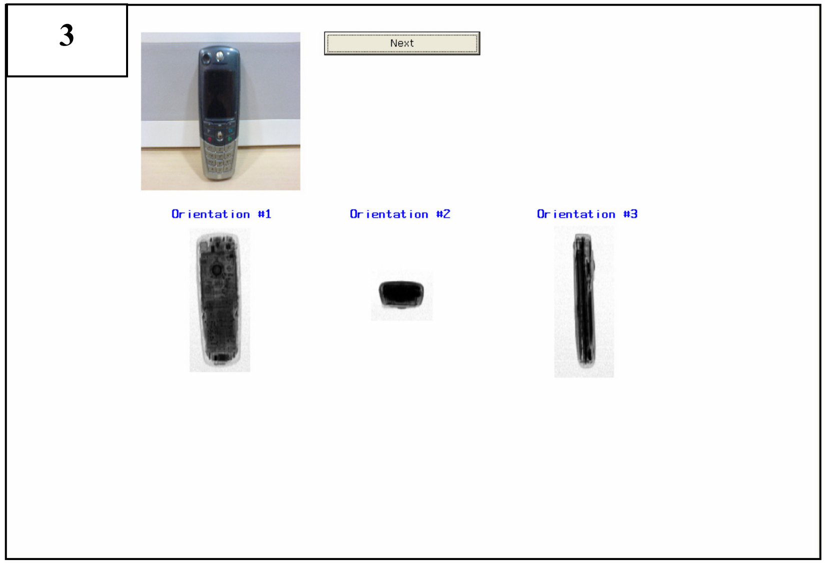
1


Next
Orientation #1
Orientation #2
Orientation #3

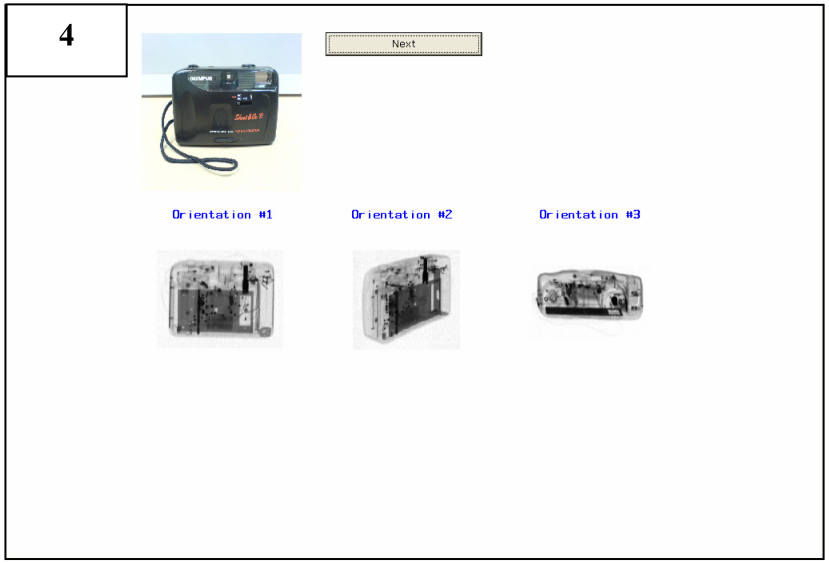
2

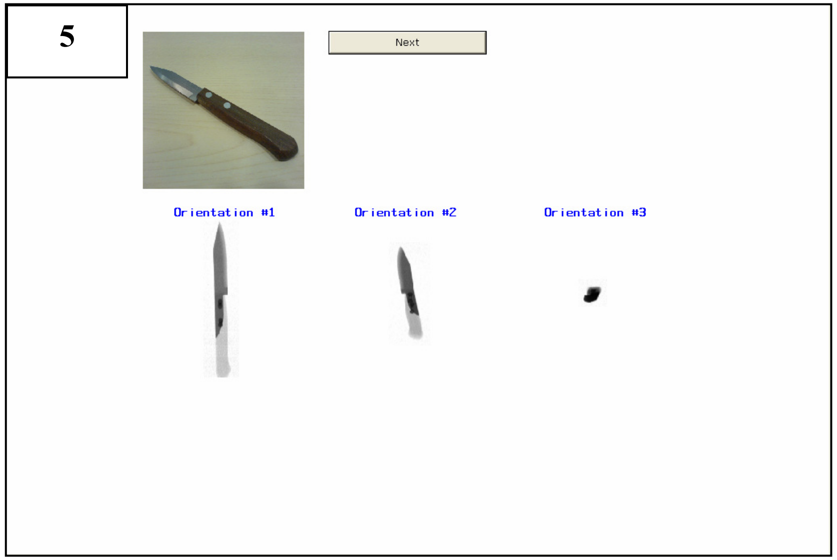

Next
Orientation #1
Orientation #2
Orientation #3

3


Next
Orientation #1
Orientation #2
Orientation #3

4


Next
Orientation #1
Orientation #2
Orientation #3



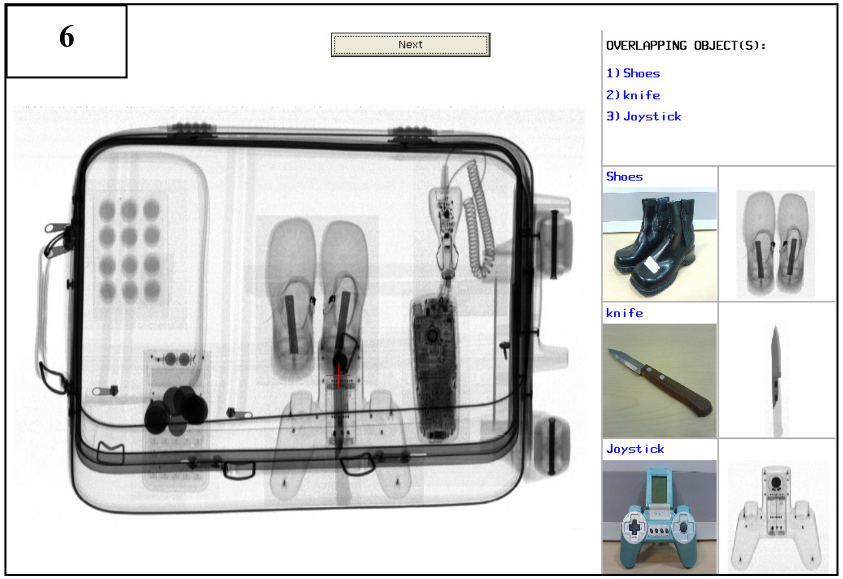


Figure 20. Stimulus familiarization training.

Note. Boxes: 1. pen; 2. stack of coins; 3. mobile phone; 4. camera; 5. knife; 6. synthesized bag with assortment of objects.

This was followed by what henceforth shall be referred to as procedural training, which was identical to that given to participants in previous experiments. Specifically, as before, participants completed a training simulation consisting of 10 practice trials with feedback in order to familiarize them with the experimental design.

Experimental Procedure

The participants were first explained the requirements of the project, and given informed consent. Following this, the author checked their visual acuity and contrast sensitivity by administering the same two short visual tests as in the previous experiment (see Appendix B). Participants' results fell within normal parameters and none was excluded as a result of these tests.

The study involved two experimental groups, each combining the two display types, Static Mono (SM) and Kinetic Mono (KM), and using two sets of images, Image Set A (A) and Image Set B (B). Half of the participants would complete SM(A)-KM(B), and other half KM(A)-SM(B); thus a counterbalanced test-retest design was employed, where the two participant groups would serve as each others' control.

It should however be noted that the targets (knives) in the set B images were deemed much more difficult to detect than the ones in Set A images, even in the KM display. They were rated as such by the author in conjunction with the three members of the supervisory team. This was an unfortunate consequence of the aforementioned technical issues (see beginning of Chapter 3) which limited the availability of new stimuli for the early experiments. Hence, it was not done intentionally, but as a result of the inability to produce new x-ray images at that stage of the project. Consequently, the experimenter was compelled to use these more "noisy" stimuli in the second part of this experiment. However, as will be seen later in the text, results indicated that despite these limitations there was still a performance advantage for the KM display over SM.

After participants were randomly assigned to one of these groups, their training began. The focus was to familiarize them with how objects appear in luggage x-ray images and more importantly how threatening objects appear in these images. Once the participants became familiarized with the procedure, they started regular experimental sessions, and their responses were recorded. Each session lasted between 20 to 30 minutes.

At the end of the experiment, participants filled out a short feedback questionnaire. They were then thanked for their participation and given their £20 compensation.

Static Experiment

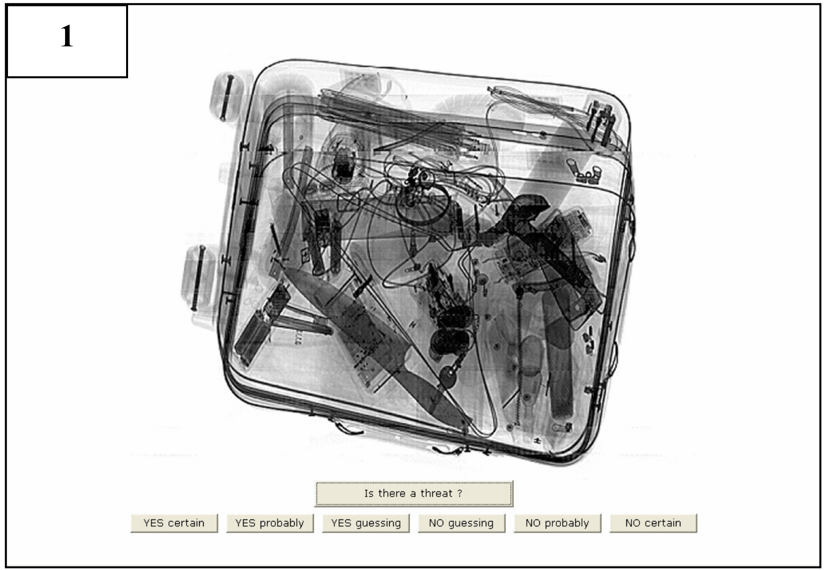
The participants were instructed to pay attention to each x-ray image that appeared on the screen, and when ready decide if the preceding image contained a threat or not. Participants were told that threatening targets were only different types of knives. They indicated their response by left clicking, with the mouse, on one of six tabs below the image, giving them the option of: ”YES Certain”, ”YES Probably”, ”YES Guessing”, ”NO Guessing”, ”NO Probably”, and ”NO Certain” (see Figure 21, boxes 1 to 3). If their answer was negative they could proceed directly to the next trial. However, if their answer was affirmative they were asked to place the mouse pointer over the object they thought was a threat and left click on it. They could then proceed to the next trial. The program was then able to recognize if the mouse click was on the target area or not and appropriately log the response.

When the new image appeared they were instructed to proceed through this, and each following trial as before. As a general rule participants were instructed to react as quickly and as accurately as possible, and their response time during each trial was recorded.

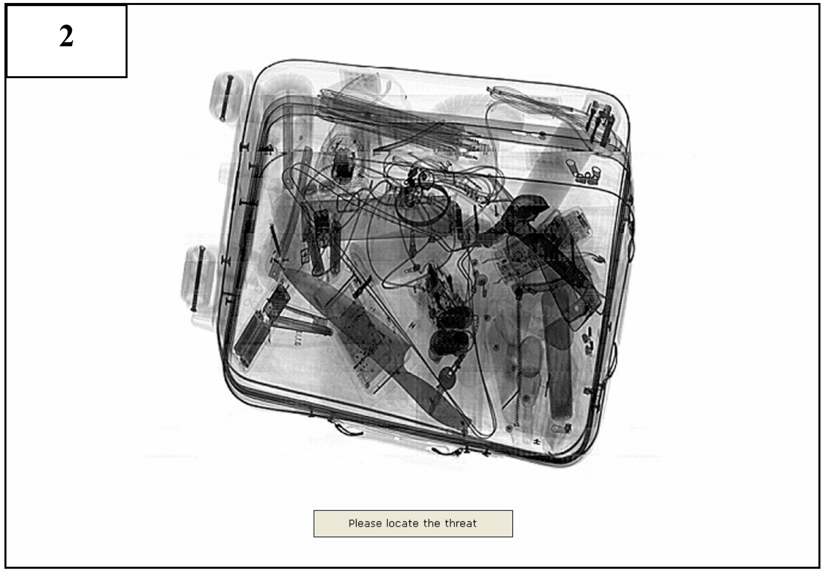
Kinetic Experiment

The procedure of this experiment was identical to the former, with the exception that all the images were presented in the Kinetic Monoscopic display. Additionally, before being able to click on an object they considered a threat, participants had to first pause the image motion using the right mouse button. They subsequently proceeded as before.

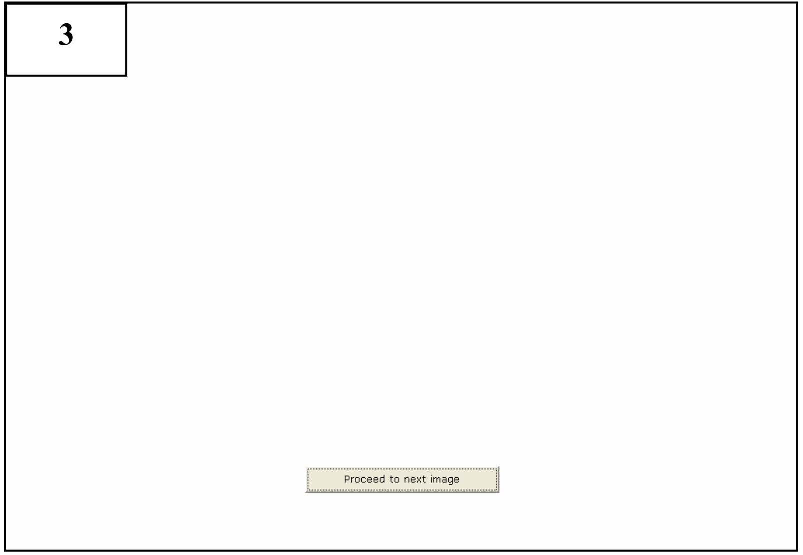
1

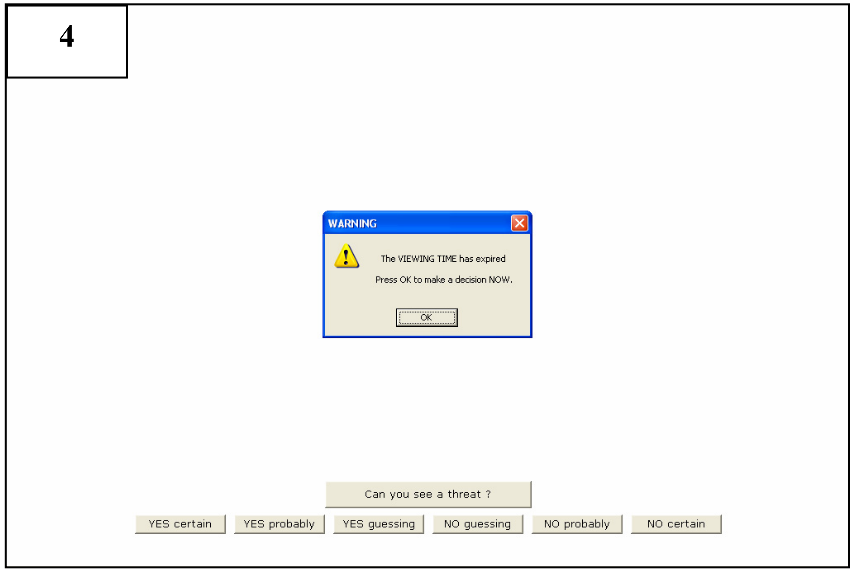

Is there a threat ?
YES certain
YES probably
YES guessing
NO guessing
NO probably
NO certain

2


Please locate the threat

3


Proceed to next image



WARNING

The VIEWING TIME has expired

Press OK to make a decision NOW.

Figure 21. Procedure for Experiments 2 (Threat Identification), and 3 (Time Limit).

Note. Boxes: 1. experiment image presented; 2. clicking on the perceived threat; 3. proceed to next trial; 4. time limit reached warning screen.

3.3.2 Results

ROC Analysis

As before, results for all participants were pooled together and using the confidence ratings data, an ROC analysis was conducted using Prism 4 (Graph Pad Software, 2005). Additionally the data were plotted in z-coordinates and linear regression was employed to obtain the slope of the lines. They resulted in a good linear fit; thus confirming the normality assumptions of the underlying distributions. However, because the slopes were all non-unit, in line with previous practice, the A_z was deemed to be the appropriate performance index for the data (Macmillan and Creelman, 2005).

The results are summarized below. Results for the two image sets (A and B) are presented in separate tables (see Table 13 and Table 14).

Table 13. *Aggregate results for image set A.*

Experiment	H	F	c	s	A_z	log (β)	RT (sec)
SM Detection	0.39	0.17	0.61	0.61	0.62	0.40	12.98
SM Identification	0.28	0.17	0.76	0.63	0.58	0.28	12.98
KM Detection	0.67	0.08	0.47	0.47	0.84	0.87	14.52
KM Identification	0.62	0.08	0.54	0.46	0.82	0.92	14.52

Note. H = the hit rate, FA = the false-alarm rate, c = the criterion, s = the zROC slope, A_z = the area under the ROC curve, log (β) = the natural logarithm of the likelihood ratio, and RT = the response time in seconds.

Table 14. *Aggregate results for image set B.*

Experiment	H	F	c	s	A_z	log (β)	RT (sec)
SM Detection	0.25	0.07	1.09	0.61	0.60	0.87	14.33
SM Identification	0.18	0.07	1.20	0.62	0.59	0.70	14.33
KM Detection	0.40	0.04	1.00	0.41	0.68	1.48	12.92
KM Identification	0.37	0.04	1.03	0.41	0.67	1.46	12.92

Note. H = the hit rate, FA = the false-alarm rate, c = the criterion, s = the zROC slope, A_z = the area under the ROC curve, log (β) = the natural logarithm of the likelihood ratio, and RT = the response time in seconds.

Results for Set A were quite robust. As can be seen in Table 13, KM resulted in a 0.22 increase in detection, and a 0.24 increase in identification, over SM. Two t-tests were performed in order to confirm statistical significance. This was found to be the case in both the detection, t(16)= -5.81, p< 0.001, and identification t(16)= -5.49, p< 0.001 components.

In the case of Set B, results were considerably less strong. However, as can be seen in Table 14, KM still resulted in a 0.08 increase in detection, as well as identification, over SM. Again, two t-tests were conducted; both the detection, t(16) = -3.45, p= 0.003, and identification, t(16)= -3.38, p= 0.004, the performance increases, albeit quite small, were found to be statistically significant.

There were small fluctuations in response times between the different display conditions. However, t-tests revealed that none of these small differences in RT were statistically significant; neither for set A t(16)= -0.758, p= 0.46, nor for set B, t(16)= 0.504, p= 0.62.

In term of response bias, decision criterion "c" and log (β) results showed that, within each image set, participants had a tendency towards giving more negative answers in the SM display, compared to the KM display. Additionally, participants tended to give more negative answers in set B, compared to set A. This is hardly surprising as the higher level of difficulty of set B images most likely prompted them to adopt a more conservative decision criterion.

There were also small differences, in each display mode, between identification and detection. For image set A the difference was greater for SM (0.04) than KM (0.02). In image set B the differences were quite small (0.01), and the same for both SM and KM.

In order to facilitate direct visual comparisons of these differences, the detection and identification standard score ROC lines (zROC) were plotted together in the same graph. However, the two image sets (A and B) are presented separately (see Figure 22 and Figure 23).

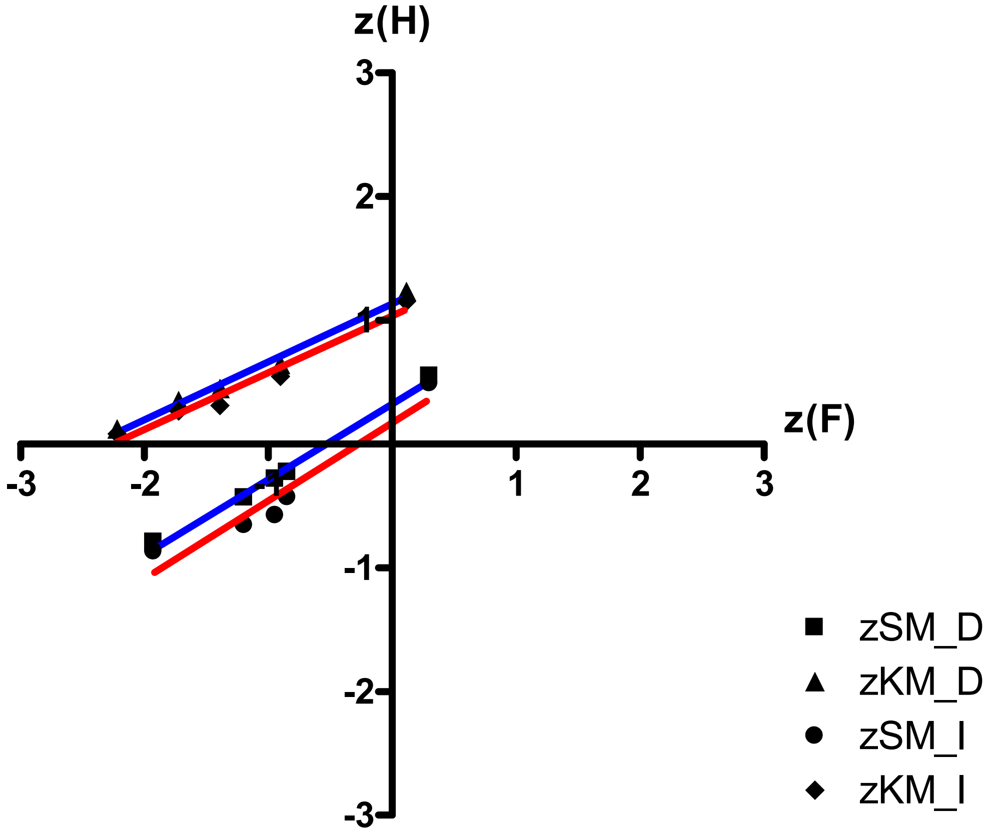


Figure 22. Image set A detection and identification zROC.

As can easily be seen from the above graph, in the case of Set A, there was a very small difference between KM detection and identification. In SM the difference between the line elevations is slightly wider. In both cases the detection and identification lines are parallel to each other.

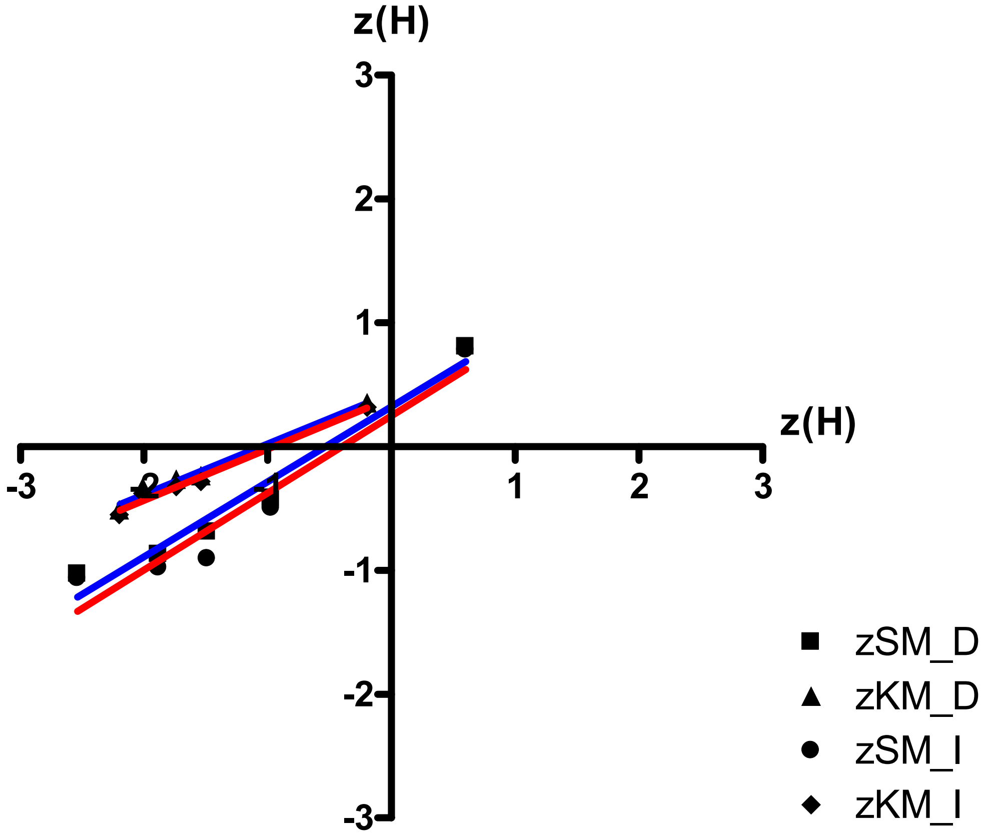


Figure 23. Image set B detection and identification zROC.

The equivalent graph for Set B reveals that there was a smaller difference between detection and identification. This is signified by the fact that in both display conditions the lines are parallel and close together.

As before, the different slope elevations of the zROC curves in the preceding figures enable the visual comparison of performance across the different conditions. In both image sets the KM display data shows increased performance, as compared to the SM display. These observations complement the A_z results presented earlier in the text.

3.3.3 Discussion

As earlier seen, in the first experiment (Vassiliades et al., 2008) it was found that KM images resulted in a superior rate of detection when compared to SM images. That was indicated by the considerable increase (0.20) of the A_z value.

In this experiment, the second of the overall series, the earlier findings were replicated, and also the range of the research extended to include the actual identification of targets. For image set A, composed of images with a moderate to severe degree of difficulty, in terms of A_z the KM display resulted in a 0.24 increase in identification performance over SM. There was a much smaller effect of KM over SM for image set B, which was composed of images with elusive targets, which likely would confound even professional screeners. This is hardly surprising given the fact that the experimental participants were inexperienced novices. However, despite this sharp increase in the level of difficulty, set B results still showed a 0.08 performance improvement for the KM display, over SM. It should be noted however that the increased difficulty of the images in set B, compared to set A, was not deliberate. This was a consequence of the technical issues affecting the experimental x-ray scanner, which prevented the collection of other better suited stimuli. Ideally the stimuli in the two image sets would have been of a similar difficulty level.

The current results support further the hypothesis that viewing images in the KM display results in superior threat detection rates, compared to the SM display. Furthermore, when taking actual target identification into account the performance difference holds up, even when confronted with very difficult targets such as those in image set B. By providing depth information to the human operators, the KDEX technique clearly helps them overcome current limiting factors such as ordinary masking, where a masking object conceals the target, and crowding, where the masking object distorts the shape of the target (Pelli, Palomares & Majaj, 2004).

In the current results, it was again observed that average response times (RT) did not appear to strongly influenced by the image display mode. In fact, the fluctuations in RT (~1.5 seconds) swung either way, and can rather be attributed to between-participant differences, and not the difference in viewing mode. Thus, despite the great performance advantage this technique has, over traditional displays, it does not appear to add any additional processing time to the task.

Further work still needs to be done, in order to expand the scope of the research to include a larger variety of threatening objects, as well as images with pseudo colour-encoding in subsequent experiments.

One of the aims of this experiment was to limit the repetition effect, caused by the completion order for the different display conditions. This was encountered both in the Pilot Study and Experiment 1. In this aspect the current experiment was largely successful, as changing the stimuli in the second part of the experiment eliminated the memory effect. This is also evident by the relative consistency in RT across experimental conditions.

The second aim of this experiment was to ascertain if there were any major differences between threat detection and identification. In other words, were participants actually seeing the threats when they said they did in the trials? The data suggest that the participants were indeed responding to the threats that were placed inside the bags, as there was only a relatively small difference in performance across the detection and the identification components of this experiment.

Specifically, examining the Az data shows that for image Set A there was a 0.04 reduction for SM identification and only a 0.02 reduction for KM identification. The differences were even more minor for Set B where there was only a 0.01 difference between detection and identification in both display types. These results validate the earlier threat detection design, which was also more straightforward in its implementation.

3.4 Experiment 3: Time Limit

The aim for the next experiment was to briefly examine what would happen if a time limit was imposed on the duration of stimulus presentation. It was decided that 10 seconds was a reasonable maximum exposure time to be able to determine if imposing such a limit there is any effect on performance.

This duration was selected partly based on the obtained RT figures from the previous experiments, and also from examining the relevant literature. For instance, Gale et al. (2000) report that experienced security screeners typically examine, each x-ray image of hand luggage, for between 6 and 10 seconds. Furthermore, in a related study, seeking to simulate effects of time pressure on screeners using inexperienced participants, Liu and Gale (2007) also limited one group of their participants to viewing x-ray scans for only 10 seconds. Additionally, this time limit is in line with what other researchers have used (e.g. Fiore, Scielzo, & Jentsch, 2004; Washburn, Taglialatela, Rice, & Smith, 2004).

The bulk of the methodology was identical to Experiment 2. Hence in the interest of expediency it will not be repeated in its entirety; the focus will instead be on the differences in the design.

3.4.1 Method

Participants

For this experiment 12 participants (5 male, 7 female) were recruited. They ranged in age from 21 to 32 (M= 26.58, SD= 3.20). Three were first time participants while the rest had participated in at least one previous experiment. At the end of the study each participant was compensated with £15 worth of gift vouchers.

Table 15. *Experiment 3 participant groups, average age, and MARS scores.*

Participants	Age		MARS Score		Display Mode
	M	SD	M	SD	
6 (2M, 4F)	26.5	3.62	1.83	0.04	Static Monoscopic Greyscale
6 (3M, 3F)	26.67	3.08	1.83	0.05	Kinetic Monoscopic Greyscale

Note. F=Female, M=Male.

Experimental Procedure

This experiment utilized the same 10 bag pairs of x-ray images as Experiment 1. Specifically, threatening objects were various knives; 1 small, 1 medium and 1 large. The threat mapping process however was identical to that in the immediately preceding experiment (Experiment 2). The instructions to the participants, as well as the training they received were also identical to Experiment 2. In short, they were instructed to inspect the bags and look for different types of knives.

The participants were separated into two equal size groups. One group viewed only SM images while the other viewed KM images. The remainder of the procedure was essentially identical to the previous experiment, with the notable exception of the introduction of the aforementioned ten second time limit (see Figure 21, Box 4). Furthermore, as each participant only completed 320 trials, they finished the entire experiment in a single day.

3.4.2 Results

Both participant groups were near identical in mean age and average log contrast sensitivity (see Table 15). In terms of response time it was noted that SM group participants spent on average 1.2 seconds longer on each trial, compared to the KM display group. However, a t-test revealed this difference was not statistically significant t(10)= 1.27, p= 0.233.

ROC Analysis

As before, the participants were first separated into two groups according to display mode (SM or KM). Then the data were pooled across participants and a standard ROC analysis was conducted. Additionally the results were plotted in z-coordinates and the line slopes obtained via linear regression. They resulted in a good linear fit; thus confirming the normality assumptions of the underlying distributions. However, as with previous experiments, the slopes were non-unit. Hence, the results are again present in terms of the A_z as the performance index. The results are summarized in the table below (Table 16).

Table 16. *Aggregate results for time limit experiment.*

Experiment	H	F	c	s	A_z	log (β)	RT (sec)
SM Detection	0.42	0.13	0.66	0.54	0.65	0.60	8.85
SM Identification	0.32	0.13	0.79	0.55	0.61	0.51	8.85
KM Detection	0.67	0.05	0.62	0.30	0.83	1.32	7.68
KM Identification	0.65	0.05	0.65	0.30	0.82	1.35	7.68

Note. H = the hit rate, FA = the false-alarm rate, c = the criterion, s = the zROC slope, A_z = the area under the ROC curve, log (β) = the natural logarithm of the likelihood ratio, and RT = the response time in seconds.

Two t-tests revealed that the differences in A_z between the two display modes (SM and KM) were significant for both the detection t(10)= -5.81, p<0.001; and identification components t(10) = -3.85, p= 0.003 of the experiment.

Furthermore a standard score ROC (zROC) plot was generated (see Figure 24). The detection and identification lines were plotted on the same graph to facilitate the visual comparisons between them.

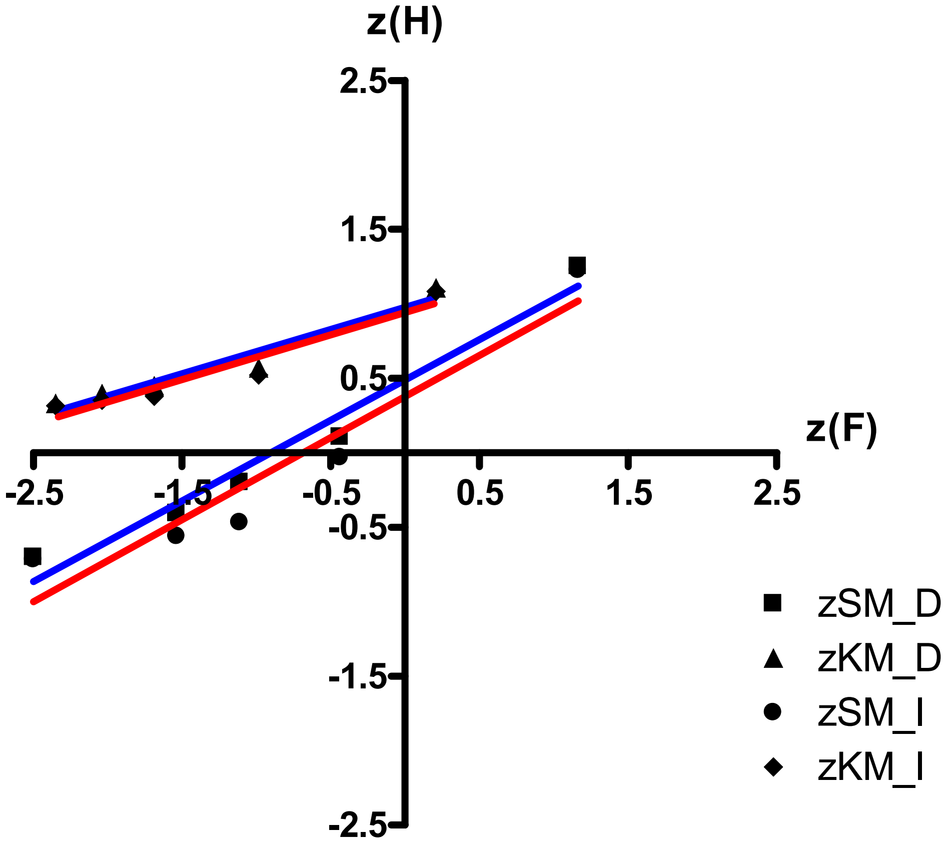


Figure 24. Time limit detection and Identification zROC

As the graph (see Figure 24) reveals there was only a slight difference in A_z (0.01) between KM detection and identification. In SM there was a somewhat bigger difference (0.04), but both could easily be accounted by experimental error.

3.4.3 Discussion

The third experiment in the series simulated a time pressure scenario where participants had a limited exposure, of 10 seconds, to the x-ray images. Overall, the data (see Table 16) reveal that performance was consistent with earlier experiments.

More specifically, there was a 0.18 improvement in performance for threat detection as measured by A_z (SM = 0.65, KM = 0.83). The performance increase was even greater for threat identification with a 0.21 increase in A_z (SM = 0.61, KM = 0.82). Once again the difference between detection and identification was quite small.

The results also mirrored the trends in the previous experiment, insofar as the decrease in sensitivity (A_z value) in the identification component, being greater for the SM condition (0.04) as compared to KM (0.01).

This is also illustrated by the relevant standardized scores ROC (zROC) graph (see Figure 24). As can be seen, performance in the KM condition is virtually identical for detection and identification. This is clearly evident by the fact that the two lines practically overlap. On the other hand the detection and identification lines are wider spaced in the SM condition; thus illustrating the wider gap in performance.

In terms of response time, even though the difference was not significant, participants in the KM group on average took less time per trial. This trend is consistent with findings from the earlier experiments, and probably has more to do with between-subject rather than between-group differences. Naturally, because of the time limit response times for both groups were reduced compared to previous experiments.

To summarize, the results did not indicate that imposing a time limit had any major effect on performance in either condition. The results were also highly consistent with what was obtained in both previous experiments (Experiment 1, and Experiment 2). Namely, an A_z increase of around 0.20 in detection performance for the KM display as compared to SM. As such it would appear that the present experiment was also successful in demonstrating the superiority of KM over SM.

Perhaps imposing an even stricter time limit (e.g. 5 seconds) would have had a more substantial impact on the results. However, 10 seconds of stimulus presentation was already the minimum time limit that could realistically be set. A shorter time period would likely not afford these novice participants enough time to confidently scan the x-ray image. Furthermore, as mentioned earlier, after careful consideration it was decided not to incorporate a time limit function in future experiments. The reasoning for this was two-fold.

First, participants generally did not like the time limit imposed upon the experiment and the disappearance of the screen seemed to serve as an aversive stimulus. Participants were thus negatively reinforced to respond before the time was up, which was not the intention.

Second, imposing a time limit may in a sense have simply added to the experimental noise. In theory, this can result in an overall loss of sensitivity to the presence of a target, and also sensitivity to its identity. Since the primary task of this project was to compare SM and KM displays, it is clear that where time limits are not severe, performance using the KM display is clearly better than with the SM display, and not merely better in the sense of statistical significance. Since screeners in the field are free to impose their own time limits, it is far more realistic in the ensuing experiments to continue to refrain from imposing any artificial time constraints. Of course for those who are concerned with the problems associated with workload and operator efficiency, as a function of time, it may well be of value to conduct studies in which time available to inspect a bag is a parameter.

3.5 Experiment 4: Pseudo-Colour

Where all the previous experiments used greyscale x-ray stimuli, this experiment employed good quality pseudo-colour images. The colours followed the industry standard coding of orange for low density “organic” materials, green for medium density “inorganic” materials and blue for “metals” (see discussion on pseudo-colour encoding in Chapter 2). As with the greyscale images black indicated that the material was opaque to x-ray radiation.

Additionally, a new type of presentation display was devised and tested. In this display a single stimulus was composed of three static x-ray images representing the two extreme angle views of each bag scan, plus the same view as that employed in the single static monoscopic (SM) stimuli of the greyscale experiments. Hence, this is labelled the Multiview Monoscopic (MV) display. The new display enabled direct testing of the effect, on observer performance, of showing different angles of the same bag without kinetic depth.

The basic research design is based on the design employed in the preceding experiments. However, some aspects of this experiment substantially differ from previous ones; those portions of the methodology will be discussed in greater length.

3.5.1 Method

Participants

Participants in the study were also recruited from the student population of the university via email invitations and poster advertisements on University notice boards. Originally 20 participants were recruited. These were randomly assigned to either the SM or KM condition. Subsequently the scope of this experiment was expanded, and an additional 8 participants were recruited, all of whom were assigned to the MV condition. Of the final 28 participants (12 female, 16 male), 4 had taken part in one of the previous experiments; the remainder were first time volunteers.

All participants were given informed consent forms to read and sign before the experiment began, and their signed copies were retained. At the conclusion of the study they were compensated for their time with £50 in the form of gift vouchers.

Table 17. *Experiment 4 participant groups, average age, and MARS scores.*

Participants	Age		MARS Score		Display Mode
	M	SD	M	SD	
10 (7M, 3F)	26.60	2.80	1.80	0.053	Static Colour Monoscopic
10 (6M, 4F)	27.90	4.86	1.82	0.034	Kinetic Colour Monoscopic
8 (3M, 5F)	24.63	4.47	1.79	0.047	Multiview Colour Monoscopic

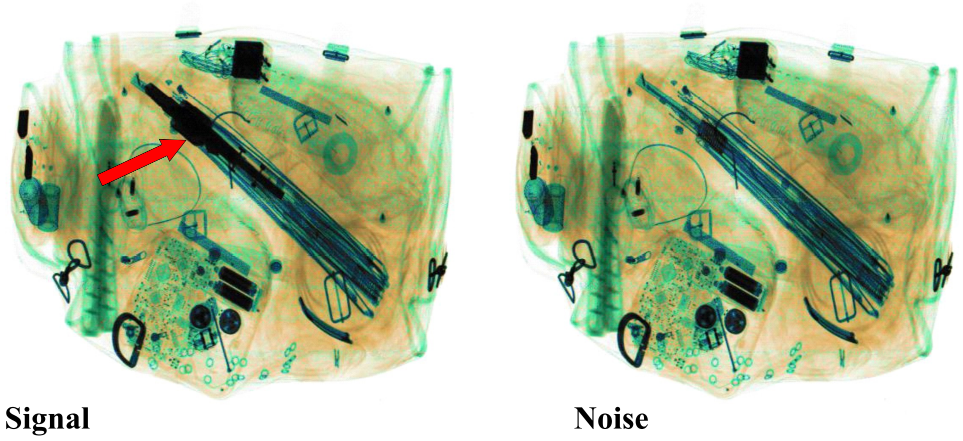
Note. F=Female, M=Male.

Four of the participants had some previous experience with the procedures, having participated in at least one previous experiment. For the remainder this was their first encounter with the research protocols. Like previous experiments all the participants were regarded as novices as they were not professional screeners and their only contact with security screening x-ray images had been from participating in these experiments.

Design

This study employed the two-response rating variant of the basic yes-no SDT design, which requires two responses from the participant during each trial. So once again after viewing the x-ray image participants were first asked if a target was present or not, and then to rate their confidence (see Figure 17).

As before, the experiments were composed of two classes of stimuli; images that contained a threatening object (signal), and images that didn't contain a threatening object (noise). An example of the two stimulus classes can be seen in Figure 25; the left image contains a revolver, which is absent in the right image.



Signal

Noise

Figure 25. Stimulus classes; signal (left) and noise (right).

For this experiment 20 pairs of threat and no threat x-ray scans were collected (see Appendix C). The stimuli will be further described later on in the text.

This image database was separated out into two smaller sets (termed Set A and Set B) of 10 bag pairs each. These were deemed to be approximately equivalent in terms of the distribution of threats in the bags, the types of threats used, and the types of bags used. Specifically, in each image set 5 bags contained a knife (or blade), 3 had a hand gun, 1 had a pair of scissors, and finally 1 bag had a grenade.

Additionally, half of the overall threats were at or near the centre of the bag, and the other half in the periphery. In terms of the actual composition of the two image sets, Set A consisted of bags 1, 3, 4, 5, 7, 9, 10, 11, 15 and 16; and Set B included bags 2, 6, 8, 12, 13, 14, 17, 18, 19 and 20. Details of all the x-ray images, as well as a short description, can be seen in Appendix C.

Within each set, the 10 different pairs of threat-present and threat-absent bags were presented in four different spatial orientations (see Figure 26). The resulting 80 images were repeated 4 times resulting in a total experimental block of 320 trials. The signal to noise ratio was again 1:1, so there were equal numbers of threat-present and threat-absent trials.

Participants again completed each experiment in 4 sessions of 80 trials. The presentation order of the images was randomized within each session. Safeguards were however put into the computer program to ensure that a threat-present and threat-absent image of the same bag would not appear sequentially.

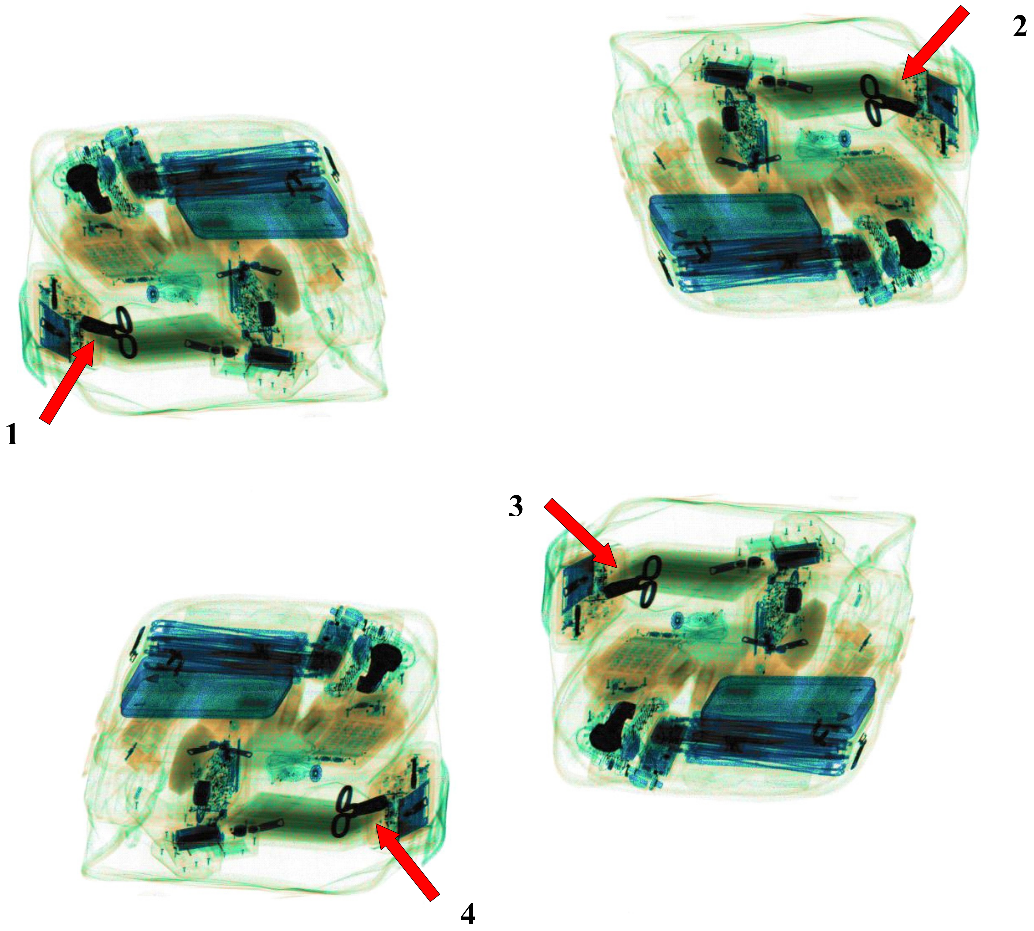


Figure 26. Original bag (top left) and three other spatial orientations.

Participants were separated into three experimental groups. Within each group they completed all trials in a single display mode (SM, KM or MV). Additionally, half of the participants viewed one set of images first (Set A), the other half viewed the second set of images first (Set B) and vice versa. Finally, since participants completed one full experimental block for each image set they each performed a total of 640 trials during the entire study.

Materials

Once again the experimental sessions were carried out in the University group's laboratory, using the same equipment setup as in the previous experiments (see Figure 13). However, for this experiment the CRT monitors were set at a higher resolution (1280x1024 pixels) and a slightly lower refresh rate (120 Hz). The illumination inside the room was again provided by reflective light from a single 100 watt incandescent bulb.

Experimental Stimuli

As with the previous experiments the x-ray images were collected using the University group's novel multiple view x-ray scanner (Evans & Hon, 2002; Hon & Evans, 2002). There was variability among bags in the amount of background complexity, the position of the targets, and the degree to which targets were overlapped by other objects in the bags. Half the targets were in or near the centre of the bag (in terms of 2D x and y coordinates), and the remaining half were in various locations in the periphery.

Additionally, the range of threatening objects was expanded. Target categories now included handguns, scissors, an assortment of knives and blades and hand grenades. Specifically, 10 bags contained a knife or blade, 6 bags a handgun, 2 bags scissors, and 2 bags a grenade (also see Table 20). For a full list and images of the stimuli used please look in Appendix C.

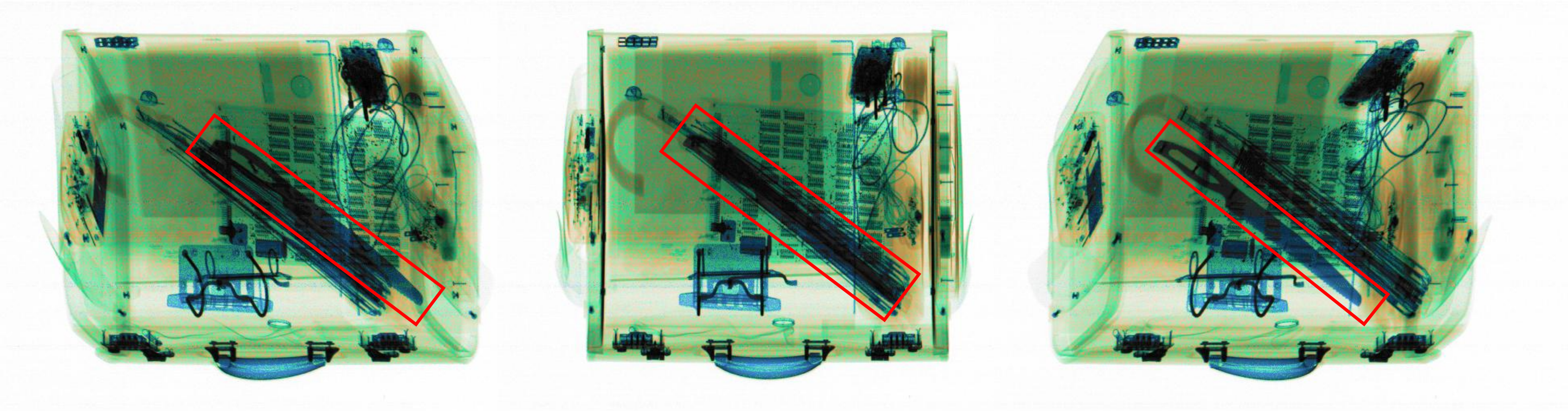
To conduct the experiments the software of the previous experiments was modified in order to make it more compatible with the improved stimuli. As before, the SM stimuli consisted of single frame x-ray images; typical of those produced by ordinary 2D airport x-ray scanners.

The KM stimuli were much improved from before. They were now composed of 41 different frames arranged into a continuous semi-rotation cycle. There was again 1° of separation between adjacent frames, so the kinetic stimuli now had a 41° span. The sequence started from the -20° frame, progressed to the +20° frame, and then returned back to the -19° frame to complete the sequence. Thus, each cycle of rotation was made up of 80 frames. The images were presented at an angular speed of 25 frames per second (or 25 degrees/second). Hence one cycle took about 3.2 seconds to complete.

This experiment also introduced a new type of presentation display, which was termed Multiview (MV). This was essentially a modified kinetic display, with the majority of the frames removed in order to eliminate the kinetic depth cues present in the KM display, but retain the angle span. Thus, these MV stimuli were composed of only 3 different frames taken from the KM sequence; the -20°, the 0° and the +20° frames. The MV sequence started from the 0° frame, jumped to the +20° frame, then back to the 0° frame, followed by the -20° frame. Each frame was presented for 2 seconds; therefore it took about 8 seconds to complete one MV cycle (4 frames presented).

A sample of the stimuli can be seen in Figure 27. In this scenario the target (large bread knife) was placed inside a briefcase, partly hidden by a large umbrella. The middle image (frame 21) served as the SM stimulus. The other two images (frames 1 and 41) show either end of the KM sequence. These three frames were also the ones used in the MV display.

The quality of the present stimuli was excellent, and a substantial improvement over the quality of the previous stimuli. Additionally, the angular range of the KM stimuli was increased almost three-fold (from 16° to 41°). However, the position of the target, relative to the bag's centre of rotation, again varied among stimuli. In the past there was concern that this, in some instances, may have constrained the angle of rotation of the target inside the bag. However, the stimuli employed here produced such as strong KDE that this would not be of any major concern.



FRAME 1

FRAME 21

FRAME 41

Figure 27. Sample pseudo-colour x-ray images, with the target outlined.

Visual Testing

As with previous experiments, before any of the volunteers participated in the experiment they had to pass three short visual tests (see Appendix B). The first two tests (Snellen visual acuity, and MARS letter contrast sensitivity) were already described earlier in the thesis so will only be mentioned briefly here. However, since pseudo-colour images were being used a colour deficiency test was also added. This enabled the experimenter to make sure that none of the participants had any issues affecting their colour perception.

First they completed the Snellen visual acuity test, followed by the MARS letter contrast sensitivity test. The third test was Ishihara's Test for Colour Deficiency (Ishihara, 1917), which is probably the most widely used tool for asessing colour vision. It consistes of a number of different plates, with an assortment of different colour dots in these plates that form the shapes of numbers (see Appendix B). People with normal colour vision can correctly see all the numbers. People with colour deficiencies see different numbers or cannot see any of the numbers at all, depending on the severity of their deficiency.

The full 21 plate version that tests for all known colour vision deficiencies was used. The test was administered at the correct distance of 75 cm, and participants had 3 seconds to give an answer for each plate.

Training

Each experiment was preceded by an extensive training period lasting about one hour. The volunteers were again given a two-part training session. First they completed the same stimulus familiarization training (see Figure 20) as before. The difference of course was that now the objects were presented in colour. Next followed the procedural training. This was improved, relative to previous experiments, and now composed of 20 practice trials (10 yes-threat, 10 no-threat), each providing the participant with feedback (see Figure 17, box 6). In addition the experimenter pointed to every single object in each x-ray image and explained to the participants the nature of the object (see Figure 28 for an example). In order to facilitate this discussion the stimuli were re-displayed after each training trial.

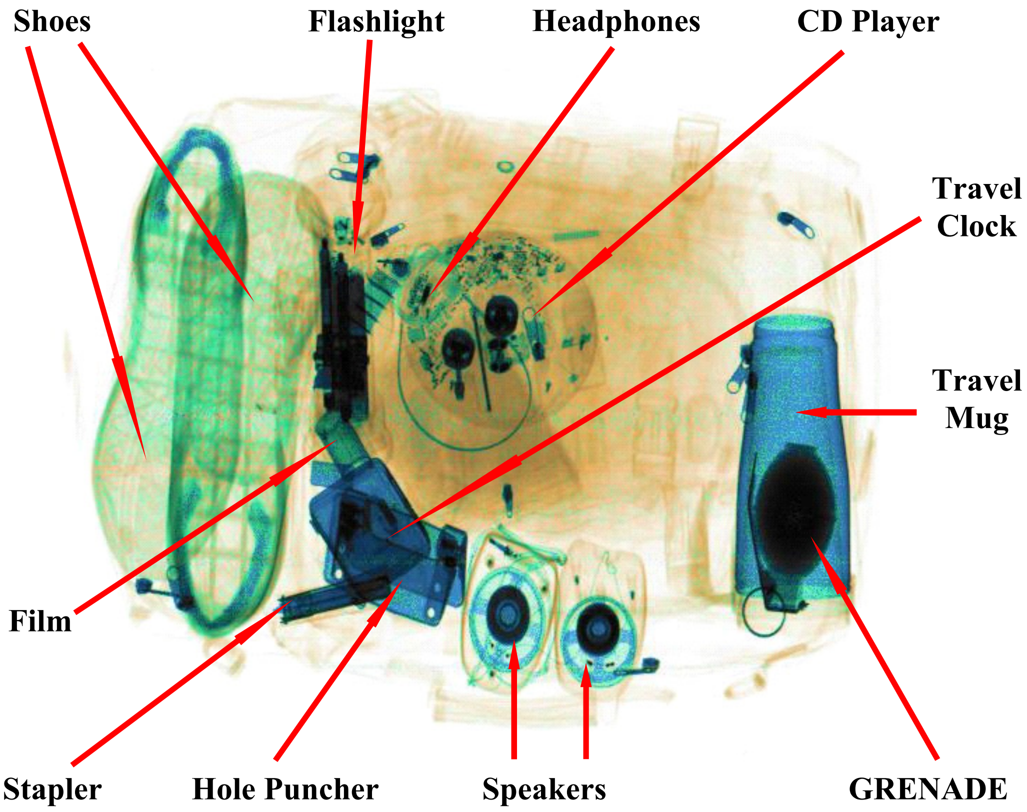


Figure 28. Training stimulus explained.

Naturally the practice stimuli were different from those presented during the experimental trials. The intention was to provide the participants with as much information as possible on pseudo-colour airport scanner x-ray images, and also how different threatening objects appear in these images. In addition the experimenter explained to the participants how they could utilize the different colour codes in the display to help them identify the types of materials in each x-ray image.

In accordance with the experimental protocol once participants had completed training they were allowed a brief break. They then continued with the experimental trials; a detailed description of the experimental procedure follows.

Experimental Procedure

As previously mentioned, participants were assigned into one of three different experimental conditions. Each group of participants viewed x-ray images in a different display mode (SM, KM or MV). Additionally, within each group the presentation order of the image sets was counterbalanced. Hence, half of the participants viewed image Set A first, and the remaining half viewed image Set B first. Overall, the experiments, in all display conditions, took place over two days; with a gap of about a week between each experiment day.

In this experiment it was decided to revert back to the more simple threat detection approach, so participants were no longer required to click on the object inside the bag they considered a threat. Consequently, the experimental procedure was near identical to the one used in the first experiment (see Figure 17).

The one important difference with this experiment however was that the participants were instructed to use the mouse to click on the response buttons on the computer screen. The experience gathered in the last two experiments indicated that participants preferred that approach over using the keyboard. Other than that the procedure during each trial was identical.

Briefly, participants were once again told to study each x-ray image carefully and once they had made a decision either click on the “Yes” tab if they thought the image contained a threat, or click on the “No” tab if they thought the image did not contain a threat. Participants were told that possible threats, present in the bags, were handguns, scissors, knives or blades, and hand grenades. After that they were asked to rate their confidence in their response. They did this by clicking on one of three tabs; “Certain”, “Probably”, or “Guessing”.

Participants then proceeded to the next trial until they completed all 4 sessions making up each experimental block. Opportunities for breaks were provided in between sessions, and on average it took them about 30-40 minutes to complete each one.

At the end of each experimental day participants also completed the NASA Task Load Index (TLX) workload assessment tool (Hart & Staveland, 1988). This is a well established scale that assesses the total amount of work that each participant put into the experiment. Participants rated the experiment on six different factors that affect workload (see Appendix B), and then rated the subjective importance of each factor to them. This is thus a weighted scale. Possible scores range between 0 (least taxing) and

100 (most taxing). The computerised version was used (NASA, 1986) which is available as a free download from the internet at the following site: http://humansystems.arc.nasa.gov/groups/TLX/

In accordance with the research protocol, at the conclusion of the study the participants were debriefed, thanked for their involvement, and paid their compensation. They also completed a short questionnaire giving feedback about the experiment.

3.5.2 Results

Visual Tests

In short, all of the participants had at least 20/25 normal or corrected binocular vision; confirmed by administering the Snellen acuity test. MARS test results were as usual within normal parameters, and the three groups were similar in their average log contrast sensitivity (see Table 17). All participants successfully passed Ishihara's test for colour deficiency.

NASA TLX

The NASA TLX scores were averaged across both experimental days for each participant. Subsequently group results were calculated which yielded mean values of SM: 57.15, KM: 51.53, and MV: 54.02. However, a one-way between-groups analysis of variance (ANOVA) test revealed that the groups did not significantly differ between them F(2, 25)= 0.765, p= 0.476. Nonetheless the trends in the data indicate that on average participants found the KM display as requiring the least amount of workload, and the SM display the most, with MV lying somewhere in the middle.

Response time

Average response time per trial was also computed (see Table 18). For SM it was 15.94 seconds, and for KM a RT of 12.38 seconds was calculated. The corresponding value for the MV display was 14.30 seconds. The one-way ANOVA test indicated that there were significant differences between the groups F(2, 25)= 4.97, p= 0.15, so additional post hoc comparisons were performed using the Tukey's HSD (Honestly Significant Difference) test. Results showed that the difference in response time was significant between the SM and KM displays, p= 0.011. However, the difference between MV and KM was not significant, p= 0.263, and neither was the difference between the SM and MV displays, p= 0.372.

ROC Analysis

Aggregate results are as follows. Data were combined across both image sets and across all participants, but separately for each display condition.

Table 18. *Data analysis for the SM, KM and MV conditions*.

Experiment	H	F	c	s	A_z	log (β)	RT (sec)
Static Mono	0.39	0.16	0.64	0.64	0.64	0.48	15.94
Kinetic Mono	0.88	0.04	0.31	0.16	0.93	0.93	12.38
Multi-view Mono	0.69	0.06	0.52	0.32	0.84	1.06	14.30

Note. H = the hit rate, FA = the false-alarm rate, c = the criterion, s = the zROC slope, A_z = the area under the ROC curve, log (β) = the natural logarithm of the likelihood ratio, and RT = the response time in seconds.

Overall performance was highest for the KM display, with the A_z reaching an overall value of 0.93, the highest value obtained during this research project. Performance on the SM display was average with an A_z value of 0.64. Finally, for the MV display the A_z was 0.84. These results reveal that the KM display offered a 0.29 improvement over SM, and a 0.09 improvement over MV. Also, the MV display offered a 0.20 increase over SM. An one-way ANOVA test showed that the three display groups were significantly different F(2, 57)= 29.88, p< 0.0001.

To evaluate between group differences additional Tukey's HSD post hoc comparisons were conducted. All the between group performance differences were found to be statistically significant. Specifically, when comparing SM to KM, p< 0.0001; SM to MV, p< 0.0001; and finally MV to KM, p= 0.038.

Additionally, the positive values of both the decision criterion "c", as well as the natural logarithm of the likelihood ratio or log (β), in all display groups, reflected a conservative propensity for participants towards making more negative decisions.

The ROC data were also plotted in z-coordinates (see Figure 29). Data from all three display conditions are presented on the same graph for ease of visual comparison.

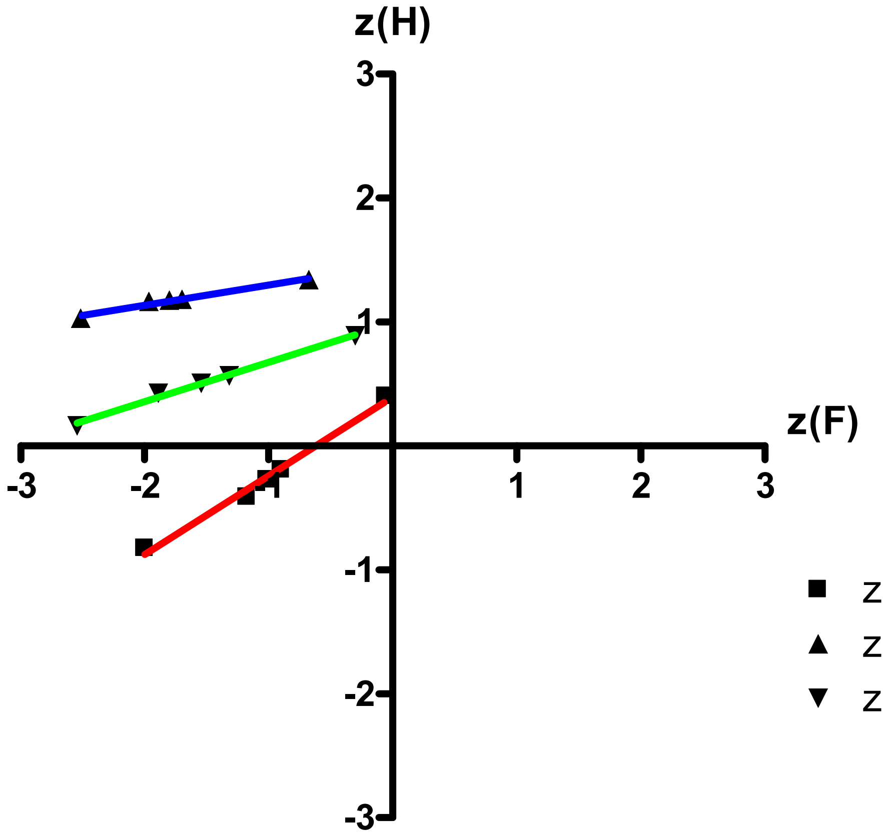


Figure 29. Standardized scores ROC of pooled data across all three experimental conditions.

The zROC plots visually represent the observations based solely on the A_z data. The KM display was clearly the highest performer with the highest elevated line. SM was the lowest performer with its line close to the chance line. MV performance was somewhere in the middle between the latter two.

The shallowness of plots also reveals a nominal inequality in the signal and noise distribution variances. Linear regression also confirmed that none of them had a slope close to 1 (see Table 18). This result supported the selection of the A_z as the appropriate performance index for the experimental data.

Threat Categories

The data from each of the 20 bag scenarios that were used in this experiment were also separately analysed. Further details of all the x-ray images used can be seen in Appendix C.

Table 19. *Individual bag results for all display types.*

	Bag Number																			
Experiment	1		2		3		4		5		6		7		8		9		10	
	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s
Static Mono	0.56	0.98	0.63	1.10	0.54	1.04	0.65	0.36	0.53	1.07	0.68	0.55	0.67	0.60	0.59	0.54	0.79	0.48	0.54	0.99
Kinetic Mono	0.89	0.08	0.90	0.13	0.91	0.23	~1.0	0.32	~1.0	0	0.82	0.25	~1.0	0	0.93	0.15	0.98	0.14	0.70	0.17
Multi-view Mono	0.84	0.21	0.69	0.46	0.95	0.28	0.96	0.04	0.92	0.40	0.74	0.58	0.82	0.50	0.75	0.39	0.94	0.38	0.56	1.64

Note. A_z = the area under the ROC curve, and s = the slope of the zROC. See Appendix C for the bag images.

	Bag Number																			
Experiment	11		12		13		14		15		16		17		18		19		20	
	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s
Static Mono	0.54	0.87	0.55	0.94	0.87	0.44	0.50	1.00	0.48	0.97	0.63	0.77	0.78	0.29	0.72	0.19	0.96	0.19	0.52	0.98
Kinetic Mono	0.99	0	0.81	0.43	0.98	0.62	0.98	0.03	0.96	0.21	0.95	0.14	0.99	0	0.98	0.08	0.99	0	0.92	0.21
Multi-view Mono	0.98	0	0.66	0.39	~1.0	0	0.95	0.19	0.55	0.52	0.64	0.36	0.98	0.20	0.97	0.02	0.99	0.63	0.84	0.27

Note. A_z = the area under the ROC curve, and s = the slope of the zROC. See Appendix C for the bag images.

As can be seen from Table 19, there was a difference in performance across different bags, as well as different display conditions. In order to make sense of all these figures it is beneficial to categorize the bags according to the type of threat placed in each.

Knives (Bags 1, 2, 3, 4, 8, 12, 14, 15, 16 and 17) were the largest threat category, with a 50% share of the targets. The second largest category were the guns (Bags 5, 7, 9, 18, 19 and 20), comprising 30% of the targets. Finally, grenades (Bags 6 and 10), and scissors (Bags 11 and 13), each had a 10% share of the total number of targets. Mean scores were computed for the three different displays' A_z results in each threat category (see Table 20).

Table 20. *Threat categories mean A_z results for each display.*

Target Category	Display mode		
	SM	KM	MV
Knives (10)	0.59	0.93	0.80
Guns (6)	0.70	0.98	0.91
Scissors (2)	0.71	0.99	0.99
Grenades (2)	0.61	0.76	0.65

Note. A_z = the area under the ROC curve.

The results indicated that there were differences in performance levels for each threat category. An extensive discussion, on the most important aspects of these differences is provided in the next section.

Furthermore, four one-way ANOVA tests were conducted to investigate the performance differences, within each threat category, between the different display types. These were followed, where appropriate, by Tukey's HSD post hoc tests. Results were significant for knives, F(2, 27)= 27.78, p< 0.001. Subsequent post hoc testing revealed that all the differences between display types were significant; specifically SM to KM, p< 0.001, SM to MV, p= 0.001, and MV to KM, p= 0.025. The differences were also significant for guns, F(2, 15)= 11.46, p= 0.001. Further post hoc testing showed the SM to KM, and the SM to MV differences were significant, p= 0.001, and p= 0.008, respectively. The difference between MV and KM was not (p= 0.549).

For the last two categories, scissors and grenades, differences were not significant; F(2, 3)= 2.84, p= 0.203, for the former, and F(2, 3)= 1.07, p= 0.445 for the latter. However this is not very surprising as, in these last two categories, the low number of bags (2) reduced the power of the statistical tests, thus increasing the possibility of a false negative result (Type II error).

3.5.3 Discussion

This was the largest and most comprehensive experiment in the course of the project. For the first time pseudo-colour x-ray images were introduced in the experiments. Additionally, the catalogue of threatening objects was expanded to include handguns, scissors and grenades.

In terms of design, this experiment used the most straightforward approach possible with an uncomplicated between-groups design. Three groups of participants were recruited, and data collected on twice as many x-ray images (20 pairs) compared with the previous experiments. The completion order, of the two different image sets, was however counter-balanced. This gave the ability to combine the data across data sets, as well as participants. Also, for the first time the data from each bag were also analysed separately. This enabled the description of the results for different target categories.

The usual SM and KM displays were complemented with a new type of display termed Multiview Monoscopic (MV). As seen earlier, creating the MV display was a direct result of one of the questions that arose during the assessment of the results from the earlier threat detection experiment (Experiment 1). It is plausible that by giving the participants a chance to view luggage images at more than one angle the contribution of kinetic depth, to the performance improvement in the KM display, could be isolated.

Thus the MV display allowed the presentation of the full angular range of the KM stimuli, but with the KDE taken out of the equation. Hence, by creating the MV display it was possible to directly compare the effect on performance of simply showing additional static frames at different angles.

Pooled Data

It is clear from the results that KM performance was exceptional. Participants in this display reached an unprecedented A_z score of 0.93, which signifies a very high overall detection performance. Performance was in fact increased for the SM display as well. This can be attributed to the much improved quality of the stimuli used in this experiment, as well as the more extensive training that the participants received.

The results are graphically illustrated by the zROC plots (see Figure 29). As can be seen by the line elevations SM was closest to chance performance whereas KM was close to perfect performance. MV performance is somewhat in the middle, but slightly closer to kinetic. The shallow slope also shows that the KM participants were more biased towards giving affirmative responses compared to SM (see log (β) results in Table 18).

NASA TLX

As can be seen by the mean NASA TLX scores, the KM display was less taxing than either SM or MV. This is an important result because it indicates that as well as offering superior detection performance, the KM display potentially reduces the burden on the operator. However it should be noted that participants' relatively midline scores (out of 100) indicate that they did not find any of the display conditions in the experiment particularly taxing. This could be explained partly by the fact that the task itself required little physical effort. Finally, although these results were obtained using novices, and not expert screeners, they still provide a useful reference point. Admittedly, a measure of workload taken from expert screeners would offer more strength, but that is something to pursue in later research.

Response Time

In terms of response time, as can be seen in Table 18, average RT per trial was overall lowest in the KM condition (12.4 seconds), and highest for the SM condition (15.9 seconds); MV lay somewhat in the middle (14.3 seconds). This is in contrast with most of the earlier experiments where the SM and KM conditions tended to have a more similar RT. However, as has been deduced from the earlier identification experiment (Experiment 2) RT mostly has to do with the attributes of the participant groups themselves; who in their entirety were novice volunteers. Hence beyond reporting these trends definite conclusions are not able to be extracted from the present RT data.

Threat Categories

The focus will now shift towards examining the results from the individual bag analysis, which were the most illuminating. It has already been demonstrated by previous experiments that the KM display offered a significant advantage over static (Vassiliades et al., 2008), but now with the inclusion of multiple types of targets, performance in different threat categories could be separately examined.

Starting with knives, the most populous threat category which comprised half of the targets, KM (A_z= 0.93) yielded a 0.34 performance increase over SM (A_z= 0.59), and a 0.13 increase over MV (A_z= 0.80). Out of the four different types of threats used knives were relatively the easiest to conceal. Even a large knife can be effectively masked by the other contents of the luggage. So even in the MV display where participants had access to 3 different bag angles they could not equate KM performance. Here the advantage that depth from motion renders to the KM display, over both SM and MV, is evident.

Performance on grenades was also relatively low for SM (A_z= 0.61), but only 0.04 higher for MV (A_z= 0.65). However in KM performance picked up a 0.15 gain (A_z= 0.76) over SM. In this case showing additional frames in the MV display did not appear to help much and there was a clear performance advantage for the KM display. Stun grenades are a difficult target and it seems that kinetic depth was once again instrumental in separating the target from the background; hence why they were more readily detected in KM.

The penultimate category, guns yielded moderate performance in SM (A_z = 0.70), and quite good in MV (A_z = 0.91). However, in the KM display performance was near perfect (A_z = 0.98). Consequently, KM had a sizeable 0.28 performance advantage over SM, but only a limited 0.07 advantage over MV. These results were not surprising for a number of reasons. As opposed to knives guns are larger and very dense metallic objects and hence inherently difficult to conceal, especially without disassembling them. This explains the relatively higher performance in the case of the MV display, and the resulting smaller advantage of KM.

Furthermore, even when well hidden inside a bag guns created a large "footprint", unless perhaps they were dismantled. Taking account of the fact that all the participants were novices it would have been extremely difficult for them to identify a dismantled handgun. However, in conclusion, it is important to note that, despite all the limitations imposed by this type of threat, out of all the displays KM still offered the best chance for detection.

Finally, for scissors moderate SM performance ($A_z = 0.71$) was again noted, but near perfect performance ($A_z = 0.99$), in both MV and KM. So the KM display boosted performance 0.28 over SM, but held no performance advantage over MV. There is a likely explanation for this result. The scissors used in the experiments were relatively large and all metal. Also scissors have a very distinctive and unambiguous shape, so they are not difficult to identify if viewed from multiple angles.

It is known from the literature that a multiple target search is less efficient than a single category search (Meneer et al, 2004), but this did not appear to be a limiting factor in the KM condition, as evident by the high performance of the participants (A_z of 0.93).

Overall the results from this experiment were very positive. It has been amply demonstrated that the KM display offers superior detection performance than SM. Additionally, with the introduction of the MV display it was also demonstrated that although showing additional angles of the same bag drastically improved performance, the KM display was superior. Furthermore, the KM display recorded faster mean response times per trial than the other two types of displays. It also relatively demanded the least amount of workload for this task. Of course as this was a laboratory based experiment further field research is ultimately necessary in order to adequately evaluate the real world workload demands of this task. Taking account of all these results it is also noted that KM is the most efficient way to present images from multiple-angle scans.

Next follows the fifth and final experiment, where the effort was to try and assess the effect of pseudo-colour in the threat detection process for the KM display.

3.6 Experiment 5: Greyscale vs. Pseudo-Colour

For the final experiment in this project the aim was to evaluate any potential performance differences between colour and greyscale kinetic displays. In the interest of expediency there was a deviation from the usual experimental protocol, and this study was conducted as a group-experiment using one of the University's computer laboratories.

3.6.1 Method

Participants

Participants were recruited by sending out mass emails and posting paper advertisements on university bulletin boards. Initially 22 volunteers were recruited, all of whom completed the experiment. However after the preliminary data analysis it was decided to exclude two of them. The reason was that the exceptionally poor quality of their data showed they had no aptitude for this task. Both of these participants were male (aged 24 and 27 years), one from each experimental group. Thus their data were excluded from all analyses reported henceforth.

Table 21. *Experiment 5 participant groups, gender and average age.*

Participants	Age		Gender		Display Mode
	M	SD	Male	Female	
10	21.7	3.50	6	4	Kinetic Mono Greyscale
10	23.6	3.95	8	2	Kinetic Mono Pseudo-Colour

Note. Data from excluded participants are not reported in this table.

Overall data are reported from 20 participants (14 male, 6 female), ranging from 19 to 31 years old. The above table (Table 21) summarises the participants' demographics.

Experimental Stimuli

This experiment used a subset of the images collected for the previous experiment (see Appendix C). Specifically, only the 10 x-ray images where the bags contained a knife or blade as a threat (Bags: 1, 2, 3, 4, 8, 12, 14, 15, 16 and 17) were used.

Furthermore, this experiment only used KM stimuli, presented in two different display conditions, either greyscale or pseudo-colour. The overall design was identical to that of the previous experiment and thus will not be repeated.

The following figure (Figure 30) contains two examples of the same bag in both pseudo-colour (left) and greyscale (right). The target, in this case a folding knife, is pointed out by the red arrows.

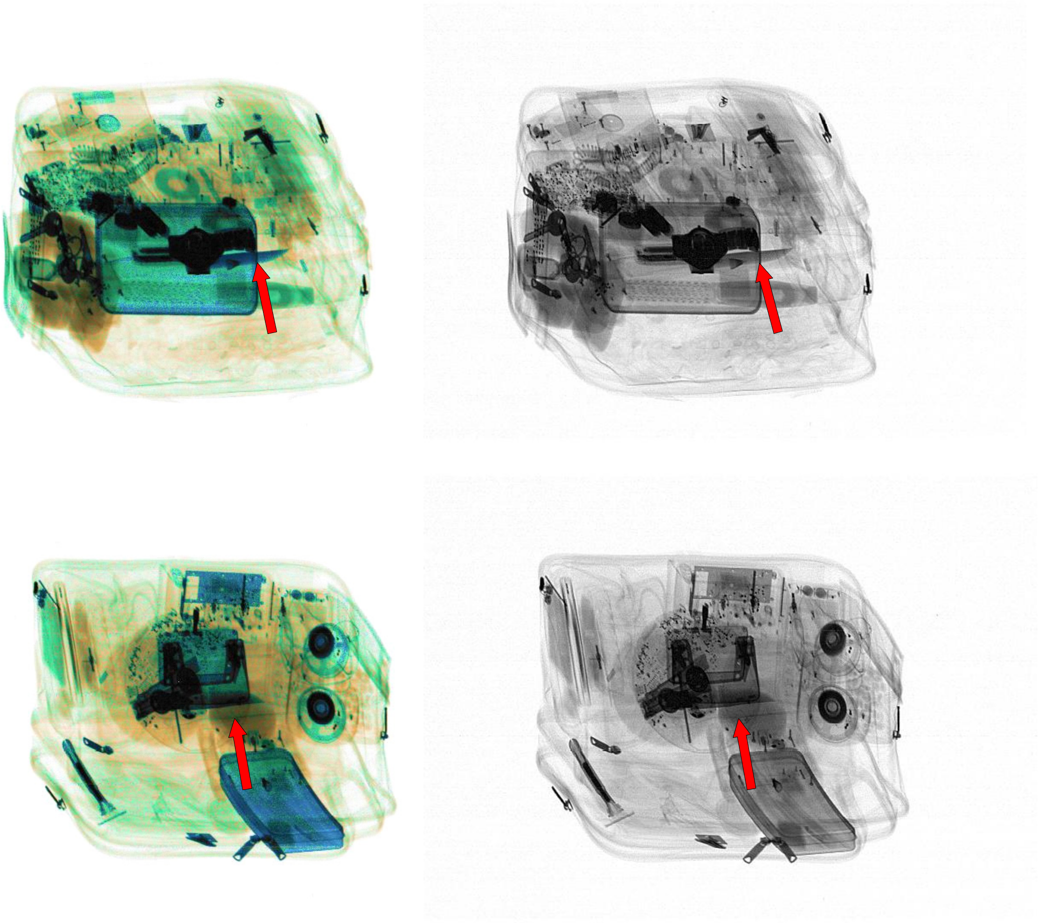


Figure 30. Sample greyscale (right) and pseudo-colour (left) stimuli.

Materials

For this experiment, in order to expedite data acquisition, the facilities of one of the university's large computer laboratories were used so all the participants could complete the experiment simultaneously. The laboratory selected was equipped with computers that closely matched the specifications of those in the University group's laboratory. Specifically the university laboratory was equipped with identical PCs, all running Windows XP and connected to 20" CRT monitors. For this experiment the monitors were set to a resolution of 1280x1024 pixels and a refresh rate of 85Hz. The software was loaded and tested on each machine prior to the commencement of the experiments.

Training

Because of concerns that the overall length of the experiment might dissuade people from signing up for the study the length of the training was slightly reduced and only offered these participants the 20 trial procedural familiarization training. Even though in the experimental trials the targets were only knives, participants were shown all the training x-ray images (including images containing handguns and the grenade). In addition, the experimenter displayed the training simulation on a large screen, using a digital projector, and explained what every single object in each x-ray image was, naturally paying special attention to the threats.

Experimental Procedure

Due to time constraints and logistical issues it was not possible to conduct the usual battery of eye tests. Instead, each participant was asked if they had any visual problems during the recruitment stage.

In order to assist with the large number of participants the author enlisted the help of a total of four additional assistants, all of whom were knowledgeable about the experimental procedures. They were briefed by the author and each was responsible for a specific section of the room. They also assisted with the training, pointed things out to participants on their individual screens, and generally made sure participants paid attention and followed the instructions of the author.

Volunteers for the study were randomly assigned to either the greyscale or the pseudo-colour condition, and all participants in each group were invited to complete the experiment together on the same day. In total three different days were required to collect data from all participants; one for the greyscale and two for the pseudo-colour condition. The reason for the additional day was the inability to obtain the full number of participants in the first instance, hence more had to be recruited and the experiment repeated.

To keep the environment between the two experimental conditions as consistent as possible all experiments were conducted in the same computer laboratory and on the same computers where possible. Additionally, all experiments were conducted on the same day of the week (Friday) and same time of day (2pm-5pm).

With the exception of the change in venue the remainder of the experimental procedure was largely identical to the previous experiments. Because this experiment involved only one set of images participants completed a block of 320 trials in the course of a single day.

As per the established procedure participants were first given informed consent and their signed copies retained. They then went on to complete the experiment in four sessions, with a break in between each session. After finishing all the trials they completed a short questionnaire about the experiment. They were then debriefed, thanked for their interest in the research, and compensated for their time with £30 in the form of gift vouchers.

3.6.2 Results

Response Time

Response times (RT) were relatively short for both groups; with per trial averages of 9.6 seconds for the KM Greyscale, and 11.9 seconds for the KM Pseudo-Colour group. A t-test revealed that the difference in RT between the two groups was statistically significant t(18)= -2.39, p= 0.028.

ROC Analysis

Aggregate results are presented in Table 22. Performance was again measured in terms of the A_Z. In the KM Greyscale display a value of 0.86 was obtained, with the corresponding value for the KM Pseudo-Colour display being 0.88. Statistical testing (t-test) revealed that the small difference in performance between the two groups was not significant, t(18)= -0.24, p= 0.82.

Table 22. *Data analysis for the KM greyscale vs. pseudo-colour experiment.*

Experiment	H	F	c	log (β)	s	A_z	RT (sec)
KM Greyscale	0.79	0.08	0.31	0.71	0.22	0.86	9.62
KM Pseudo-Colour	0.81	0.07	0.31	0.73	0.29	0.88	11.86

Note. H = the hit rate, FA = the false-alarm rate, c = the criterion, s = the zROC slope, A_z = the area under the ROC curve, log (β) = the natural logarithm of the likelihood ratio, and RT = the response time in seconds.

In the following figure (Figure 31) the pooled ROC data were also plotted in z-coordinates, thus graphically illustrating the analysis results.

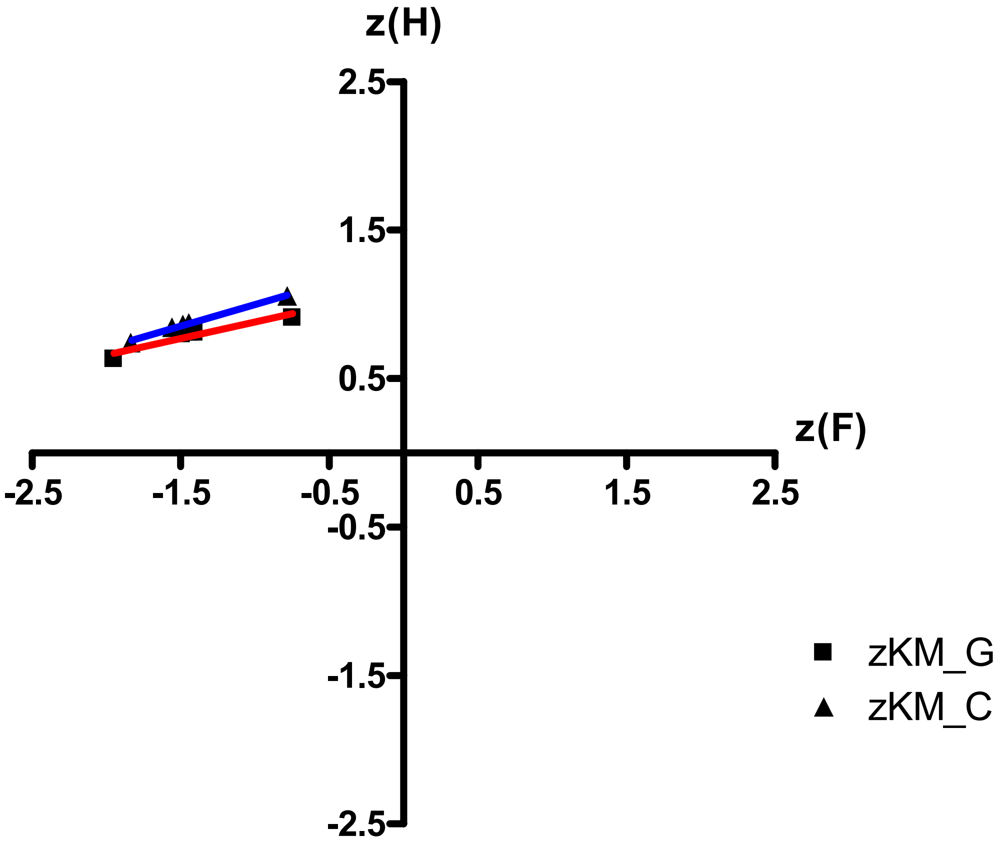


Figure 31. Standardized scores ROC of pooled data across both experimental conditions.

Table 23. *Individual bag results for greyscale vs. pseudo-colour KM displays*.

	Bag Number																			
Experiment	1		2		3		4		8		12		14		15		16		17	
	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s	A_z	s
KM Greyscale	0.92	0.08	0.87	0.13	0.79	0.23	0.99	0	0.85	0.66	0.60	0.48	0.93	0.14	0.89	0.56	0.84	0.09	0.97	0.59
KM Pseudo-Colour	0.92	0.19	0.81	0.30	0.90	0.46	0.99	0.10	0.77	0.39	0.71	0.37	0.91	0.15	0.97	0.11	0.81	0.16	0.98	0.53

Note. A_z = the area under the ROC curve, and s = the slope of the zROC. See Appendix C for the bag images.

Individual Bag Analysis

As with the previous experiment, a separate ROC analysis was conducted on the data from each bag type used (see Table 23). As the bags used here were a subset of the bags collected for the previous experiment (see Appendix C) the same bag numbers were retained. In this manner the data can be easily related between the two experiments.

As can be seen in the preceding data table (see Table 23) for some bags there was little or no difference in A_Z between pseudo-colour and greyscale, for others a clear advantage for pseudo-colour and for the remaining greyscale seemed to offer better opportunity for threat detection. However, individual bag results between the two groups are significantly correlated r = 0.80, p= 0.006. A more thorough discussion of these results follows in the next section.

3.6.3 Discussion

This being the last experiment conducted during the course of this project, it was possible to draw on past experience and conduct this as a group study instead of the previously established procedure of one or two people per time sessions.

This approach of course had its limitations. For instance, as was briefly mentioned in the previous section, because of the time pressures and the logistics involved it was not feasible to conduct the usual battery of eye tests on each participant. Instead they were individually questioned about their vision, and from their responses, along with the experimenter's experience, their eligibility to participate was determined.

However, despite using such subjective criteria, given the demographic group that the participants came from (university computing department students) no issues were anticipated, with regards to their ability to observe these images on a CRT display; a task that they routinely perform on a daily basis as part of their coursework.

Pooled Data

Overall results show that participants performed quite well in terms of threat detection, with high A_z values of 0.86 and 0.88 respectively. Thus, the KM Pseudo-Colour display yielded 0.02 higher results than the KM Greyscale display, a mild increase at best. Decision criterion, as well as log (β), results also show that the two groups were almost equal in their response bias. Both indexes resulted in positive figures; hence participants were more biased towards giving negative responses (conservative criterion).

Additionally, by examining the zROC graph of the pooled results (Figure 31) the similarity in performance between the two groups can also be observed, as the two zROC lines are quite close together. The same plot also indicates that participants generally felt quite confident in their responses. This is evident by the close clustering of the data points on both plots. This is not surprising as both groups viewed images in the KM display, and all the targets were from the same threat category.

However, in terms of response time, participants in the KM Pseudo-Colour group spent an average of 2.24 seconds longer on each trial than in KM Greyscale. As these were novice participants, it is theorized that their unfamiliarity with pseudo-colour images probably caused them to be more thorough during their visual search for targets.

Individual Bags

Overall, the results were as anticipated based on the results of earlier work. As can be seen in Table 23, although overall performance was largely identical between the pseudo-colour and the greyscale display, when looking at individual bags in some cases there are differences; while in others there are not (as was the case for the different display types in the immediately previous experiment).

A close examination of these scenarios will now follow and once more attempt to ascertain the reasons behind these differences. This step is crucial as it can inform on the effective use of colour in KM displays.

It is clear that for some bags pseudo-colour, or the lack of it, undeniably played a role in the likelihood of target detection. For instance, for bags 3 and 12 (which incidentally are the bags in Figure 30) pseudo-colour yielded a 0.11 improvement over greyscale in both instances. This result was also true for bag 15 where there was a 0.08 improvement. Finally, bag 17 demonstrated a 0.01 increase in the value of A_z. The most likely explanation, in the case of the first three bags (bags 3, 12 and 15) is that the colour-encoding of the knife's blade made it stand out better from the background in the pseudo-colour condition, so it was more likely to be detected by the participants (see Figure 30).

However in certain instances there was an advantage for greyscale over the pseudo-colour display. This can be seen in bags 2 and 8, where greyscale actually yielded better performance results. The increase in A_z values was 0.06 for the former, and 0.08 for the latter. It is theorized that the intensity of the orange background colour may have actually hindered detection in the pseudo colour display (see Figure 32). Although without looking at the actual KM stimuli, the reader may have trouble visualising this.

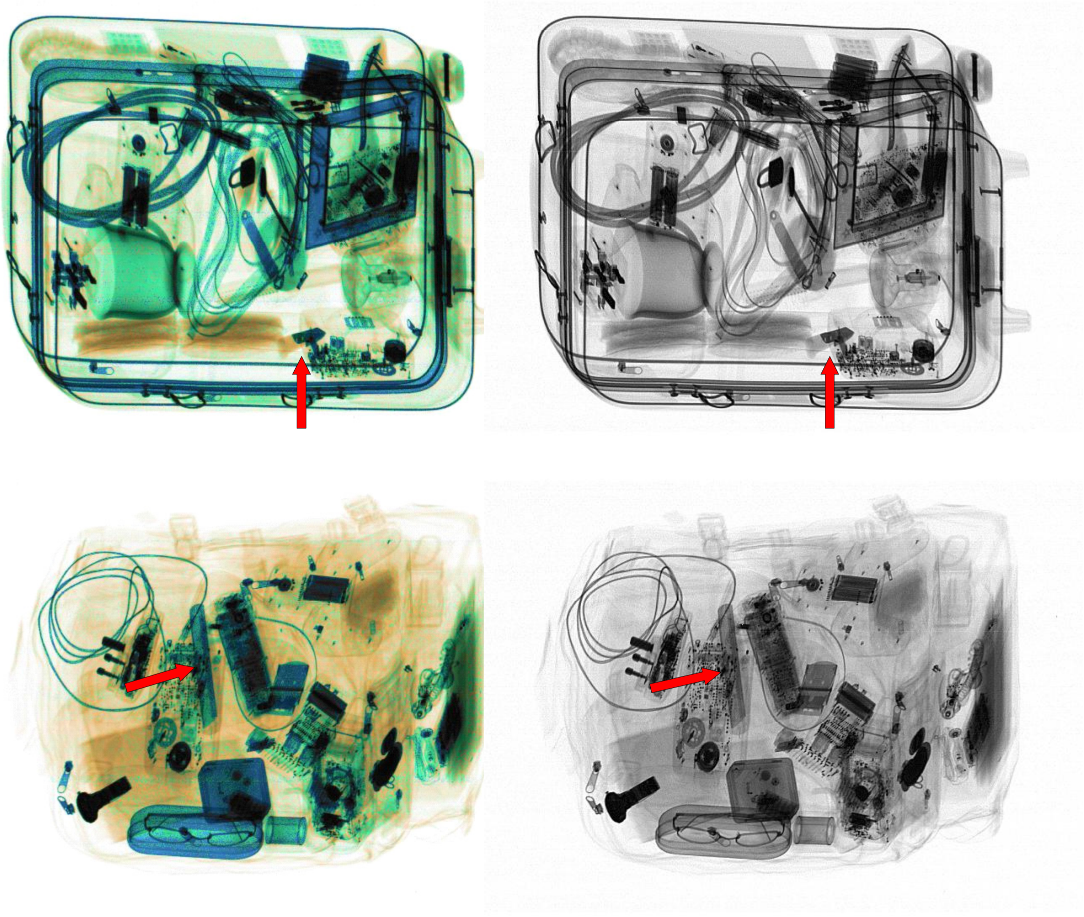


Figure 32. Stimuli where greyscale held an advantage over pseudo-colour.

Along similar trends, bags 14 and 16 showed a smaller greyscale advantage of 0.02 and 0.03 respectively. Perhaps also in these bags the orange colour of the background affected, but to a lesser degree, the visibility of the target. Thus, resulting in a somewhat reduced performance for the pseudo-colour display. However, the differences are so slight that in these two cases the possibility that this may be due to within-subject differences cannot be discounted.

The remaining bags (1 and 4) showed no difference in A_z between pseudo-colour and greyscale. For these bags it is likely that the shape of the threatening object was sufficient for the observer to detect it and there was no difference in detection between the colour and stance.

It is likely that attributes of the stimuli account for the major differences between the two displays. However, as aforementioned, some of the smaller performance differences (e.g. Bags 14 and 16) could be due to differences between participants in the two groups rather than in the images themselves. This is also likely the case for bag 17, where any likely reasons for an advantage in either display type could not be identified.

This experiment is by no means definitive and many more factors, such as the relative luminances of the image, could be at play here. One important consideration is the inexperience of the participants, which may have prevented them from using the available colour information effectively. However, a discussion on how to best make use of the information this experiment provides, with respect to how to best apply a colour pallet in future KM displays, follows in the next chapter.

4 Chapter Four: Conclusions

4.1 Overview and Discussion

The preceding empirical work will now be summarized and the lessons learned with each successive experiment examined. Additionally, later on there will be a discussion on the strengths and limitations, with the focus on drawing final conclusions from this project as a whole.

As discussed in the previous chapter, this investigation began with two exploratory pilot studies. Results from both experiments were very encouraging. In both experiments a performance advantage of the KM over the SM display was found. The pilot experiments made it clear that in future experiments significant differences in variance between the hypothetical distributions underlying the hit and false-alarm data would likely be encountered. Hence, real advantages could accrue by employing the Yes-No detection design, as this allows the efficient use of more ratings than does the 2AFC design; thus making it possible to plot ROC curves in greater detail. In addition, a Yes-No detection design appeared to more closely resemble the real life security screening scenario that was being simulated under laboratory conditions. Hence the 2AFC discrimination approach was put on hold in favour of the single interval Yes-No rating design.

Overall the pilot study provided an important stepping stone. Testing different designs and evaluating the results helped set the foundation for the later work on the project.

The first of the main experiments that followed was an improved version of the Yes-No pilot experiment. The results robustly demonstrated the advantage of the KM display over SM. Additionally the query about the origin of the memory effects biasing some of the data was resolved. The conclusion that repetition effects were at fault was confidently reached. Specifically, these repetition effects occurred in the case where participants viewed images in the KM display, followed by the same images in the SM display.

Being able to directly assess if participants were actually identifying the targets placed inside the bags was the primary goal of the second experiment. The results confirmed that detection and identification performances were close enough to each other to validate the use of the less complicated detection design. Additionally by employing different image sets in a test-retest design it was demonstrated that repetition effects were limited to the case where the same images were used. This was however confounded by the increased difficulty of the second image set. This might seem like an obvious conclusion at first, but bearing in mind reports of memory for the KDE, this issue deserved a more detailed investigation. It should be stated that no evidence was found of a memory effect similar to that described by Wallach, O'Connell and Neisser (1953) in their study of the KDE. Of course this design does not entirely rule out the possibility of other practice effects, as the participants' familiarity with the task progressively increases.

The smaller third experiment allowed testing of the notion of setting a limit on the stimulus viewing period. Although results from the time limit experiment were interesting, implementing a time restriction function in the design of future experiments was decided against. The reason for this was the frustration it caused the novice participants, and additionally because the time constraints in real-world scenarios are more of a self imposed nature, as the screener tries to clear the backlog of bags that need to be inspected.

The introduction of pseudo-colour, in the fourth experiment, was the defining moment in the project. This was the largest and most comprehensive experiment in the course of the project, both in terms of participants, and of the number of distinct stimuli used. Another important innovation was the introduction of four separate threat categories: knives, handguns, scissors and grenades. This on the one hand complicated the detection task for the participants; but on the other hand it demonstrated that the KM display was reliably better than SM for many different types of targets.

Furthermore, demonstrating the superiority of KM against the MV display greatly improved the validity of the results. Although, as expected, showing 3 different static angles in the MV display significantly improved performance, over SM, the KM display was superior overall. Thus, this experiment directly demonstrated the role that depth cues play in the performance improvement afforded by the KM display. In terms of different target categories, the KM display was superior for all types of threats used (knives, guns and grenades) except scissors, where performance was the same as MV. This was not very surprising given the nature of those specific targets (see the discussion in Chapter 3: Experiment 4).

In conclusion, as expected presenting different perspectives, even in static form, enhanced detection in the MV display simply because of the availability of more information. This is because 3 different views provide more information than mere repetition of the same view 3 times can. However, this type of presentation is limited by the capacity of the observer to process large amounts of information. The coherence provided by the KM display allows the observer to assimilate more information in a similar time period. Furthermore, the kinetic display enables the extraction of relevant depth information, and provides phenomenally smooth motion across a large number of viewpoints.

In the fifth and final experiment, performance on the new KM Pseudo-Colour images was contrasted with same quality and resolution KM Greyscale images. This experiment used a single threat category, knives. The logic for this was two-fold. On the one hand, given the group test design of this experiment, it was far less complicated to train participants in a single threat category. Second, knives are generally constructed by a variety of materials (e.g. wooden handle and a metal blade). Thus, they appear segmented in different colours, making them the most challenging type of target. Hence any performance differences between pseudo-colour and greyscale would more readily surface.

The most important contribution of this experiment was the verification of the suspicions that, under certain conditions, the colour-encoding of some x-ray images might possibly hinder detection. However the more important finding was that colour can greatly assist detection in other instances. It should be noted though that professional screeners operating in the field have the capability to switch between a greyscale and pseudo colour display for any particular scan. Sometimes, by employing a dual monitor setup, they may even view them simultaneously (Gale et al, 2000).

Of course these results have to be treated with a degree of caution. For instance, pseudo-colour enhancement of greyscale images has been shown to aid performance of professional screeners viewing 2D x-ray images (cf. Abidi, Zheng, Gribok & Abidi, 2005). However, the present experiment used inexperienced participants, who were not intensively trained on how to optimally interpret pseudo colour images. As such it was limited in scope, and definitive conclusions cannot be drawn from such a brief study. This experiment does however demonstrate the need for further research into this topic.

In terms of response time, even though results varied from experiment to experiment there was a consistent trend of kinetic results being similar or lower than those for static. Thus, the only conclusion that can be safely reached, in the context of this research, is that the KM display did not significantly increase participant's response times. This is despite providing significantly more information for the participants to process when making a decision if a threat is present in a bag or not. Hence, taking into account the performance increase provided by the KM display, there was no compromise between speed and detection accuracy in this laboratory setting.

However it should also be noted that the average period participants spent viewing each x-ray image, regardless of display mode, was itself quite long when compared with those expected of real world screeners. As discussed earlier (see Chapter 3), experienced security screeners typically examine, each x-ray image of hand luggage, for between 6 and 10 seconds (Gale et al., 2000). However, in most of the present experiments average RT was over 10 seconds. But this most likely reflects the inexperience of the experimental participants.

In terms of response bias, in all experiments the values of the decision criterion "c" were positive. This was a result of the miss rate (false negative) always being higher than the false alarm rate (Macmillan & Creelman, 2005). This tendency was also reflected in the log (β) values, which were consistently positive as well (Wickens, 2002). There were of course variations across different experiments and conditions; however as a general observation the constantly positive criterion and log (β) values indicate that participants tended to be more biased towards giving negative responses.

This is understandable as their unfamiliarity with x-ray images most likely prompted them to generally adopt a more conservative criterion in responding to all stimuli during experiments.

Overall, all the data collected during the course of this project, consistently render support to the notion of a threat detection performance advantage for kinetic displays. It was also demonstrated that KM is overall superior regardless of the type of threat present, and it does so within the same time frame and without being more taxing than existing technology.

4.2 Laboratory Experiments vs. the Real World

Another aspect of this research that is worth discussing is how, as a function of its laboratory setting, it differed from the real world. One obvious difference is the environment. Professional screeners often have to operate in a busy and noisy airport, and furthermore share the same environment with several of their colleagues (Liu, Gale & Song, 2007). Consequently there are several potential environmental distractors that can influence their attention to the task at hand. Furthermore, it is also possible that work procedures may differ across different settings (e.g. across different airports). These extraneous variables are inherently controlled for in laboratory experiments, as the experimenter typically maintains a high degree of control to what participants experience during the procedure.

However, even in the case of field research, in order to conduct experiments and collect meaningful data sometimes some degree of deviation from the real-life task is unavoidable. For instance, if the actual mechanics of the real world security screening task are taken into consideration then one immediate problem is that it does not easily lend itself to the accurate recording of the responses and response times. This is because it is a continuous task, and not composed of a series of discrete trials as in laboratory experiments. Thus even in field experiments, which unarguably possess the highest degree of realism, some modification of the actual task procedure would be necessary in order to collect meaningful response time data.

Another difference is the low probability of targets in the actual screening task. This has been discussed earlier (see Chapter 2). Research suggests that the low probability of targets in visual inspection tasks can negatively affect vigilance and consequently may lead to miss errors, (Wolfe et al., 2005), although there is as yet no conclusive consensus on how to effectively remedy this particular issue (cf. Wolfe et al., 2007; Fleck and Mitroff, 2007). However, screener vigilance, or the lack thereof, was not a consideration in this investigation. The issue of concern was potential miss errors due to the inherent lack of depth cues in static (2D) x-ray images. Hence detection performance in the two competing displays needed to be compared under controlled conditions, ideally achieved in a laboratory setting. Additionally, even in a field experiment, affording the highest degree of realism, knowledge on the part of the participants that they are being evaluated can potentially influence their degree of vigilance during the task.

Furthermore, it seems likely that differences in the probability of occurrence of targets would affect screener bias (cf. Singer & Wixted, 2006, for a discussion on factors that influence criterion shifts). In point of fact, adjusting the target prevalence in SDT experiments is known to influence the location of the decision criterion, without affecting sensitivity per se (cf. Wolfe et al., 2007). Hence, maintaining equal numbers of signal trials and noise trials (1:1 signal to noise ratio) is tantamount to keeping the decision criterion constant; with respect to vigilance. This allows the more accurate measurement of sensitivity differences, in the parameters of interest that are relevant to the present investigation; such as target type (knife vs. gun), display type (kinetic vs. static), or use of pseudo-colour vs. greyscale.

To reiterate, while it is evident from the literature that low target prevalence does have an effect on vigilance, and consequently detection performance, this was not relevant to the purposes of the present investigation. This however could be looked at as part of future research.

However, it should be noted that important research into more effective screener training regimes is underway. For instance, by applying medical image inspection modelling methods Gale et al, (2005) are developing techniques, whereby using eye tracking, the specific causes of missing a target can be identified. This enables the categorical classification of these failures (search, detection, or interpretation errors), and thus their respective causes can be pinpointed and consequently addressed by additional screener training. Furthermore, in a large study, involving 97 professional security screeners, Michel et al. (2007) utilized adaptive computer based training. Results suggest that the training led to an increase in detection performance, as well as faster response times, when a threatening object was present in a luggage x-ray scan.

In conclusion, in a laboratory setting all these aforementioned ancillary aspects of world tasks are systematically factored out in order to purely concentrate on the main question of interest. In this case the query quite simply is if, under equal conditions, observers are better at finding the prescribed targets when viewing kinetic images, as compared to static. Naturally, it is always possible that some extraneous variables can creep in, as a result of experimental error (e.g. participants not following the experimenter's instructions), and influence the results of laboratory experiments. Hence, care should always be taken to maintain tight experimental control in laboratory settings; as was the case for this research.

4.3 Accomplishments

The main and most important accomplishment of this research project is having successfully answered the research questions put forth at the outset of this thesis. Simply and succinctly the consistent superiority of the kinetic displays over the static ones was amply demonstrated.

This is a very important finding as it validates the entire research effort behind the inception, design and implementation of the KDEX imaging technique. It furthermore laid the foundation for justifying the construction of a commercial prototype with the aim of conducting field trials in the near future.

To that effect the main sponsor of this research project, the United States Department of Homeland Security, has awarded contracts to two USA based companies to independently develop KDEX prototypes and conduct field trials. This was a direct result of the research presented in this thesis.

4.4 Limitations

Unfortunately, as is often the case in research this project was not spared certain setbacks. As extensively discussed earlier (see Chapter 3) this imposed severe limitations on the availability, and quality, of stimuli for the beginning stages of this investigation. In order to resolve these issues the unique imaging chain of the x-ray scanner had to be rebuilt during the course of the project. Consequently, these initial technical issues were overcome and high quality pseudo-colour x-ray images were collected.

Another issue was the type of participants employed. As is often the case in research a convenience sample, taken primarily from the University's student and staff population was relied upon. Hence all the participants, even repeat ones, have to be classified as novices.

Consequently, training these novice participants in order for them to acquire even the fundamental principles of security x-ray screening was a time consuming affair. This process occupied almost one third of total experimental time. Additionally, it required a significant effort on the part of the author to both maintain a high training standard, and at the same time be consistent across the multitude of participants.

On the other hand, the fact that good quality results were obtained, despite using novice volunteers, further demonstrates the strength of the KDEX technique.

4.5 Future Research

There are several directions for future research to follow. For a start, a more comprehensive investigation of the advantages and disadvantages, under different circumstances, of pseudo-colour KDEX imagery needs to be conducted.

As the last experiment indicated under certain conditions colour aided the detection of threats but in other instances it hindered it and most of the time made no difference.

One must bear in mind that the colour-coding industry standards used here were developed with static displays in mind. Further research needs to be undertaken to identify the most suitable colours for use in kinetic displays.

The effectiveness of the current industry standard colour-coding scheme has been the subject of discussion since its introduction. Some work has already taken place in an effort to assess if shifting the colour-coding pallet of dual energy x-ray images can help increase detection performance (cf. Hilscher, 2005; Kase, 2002).

For example, Abidi et al. (2005) presented screeners x-ray images encoded with several different pseudo-colour methods. They found that images utilizing color mapping schemes, designed based on human perceptive and cognitive features, outperformed ordinary greyscale images, in terms of target detection. Additionally these schemes were rated higher by the screeners themselves.

Additional evidence on the contribution of the roles of different colour combinations in figure-ground relationships is found by examining the literature on scene recognition (cf. Toet & Franken, 2003 for a discussion on different image fusion schemes).

These other experiments in the literature have only been conducted with two-dimensional stimuli in mind. The present experiments used three-dimensional stimuli, but so far only compared the current industry standard colour scheme to greyscale. Clearly there is a gap in the literature, and further research needs to take place to specifically determine the optimum colour codes that should be used to complement the images produced using the KDEX technique.

One potential avenue worth exploring is borrowing pseudo-colour encoding schemes from medical imaging (cf. Parker, Zhang & Rubens, 2006, for a general discussion on colour in medical imaging). Of particular interest would of course the relevant palettes used in 3D medical techniques, such as computerized axial tomography (CAT), and Magnetic Resonance Imaging (MRI). The aim is of course to incorporate some of the lessons learned from that field into improving the colour encoding in future kinetic stimuli.

Another potential research path involves integrating kinetic depth with stereoscopic displays. As previously stated (see Chapter 2), the spontaneous reversal of the perceived depth plane in KDE displays is a frequent occurrence. This issue with kinetic depth perception did not escape the attention of researchers, who early on after the discovery of KDE began investigating it (Day & Power, 1965). Several studies confirm that the introduction of additional depth cues (cf. Fang & He, 2004), such as disparity, in these types of stimuli could effectively address this problem by determining a joint depth map (Sperling, & Dosher, 1994).

Additionally, Schwartz & Sperling (1983) caution as to the effective usage of luminance contrast in order to disambiguate depth in computer-generated KDE displays. This is important because observers are inclined to perceive high-contrast objects as being closer than low-contrast objects (Sperling, & Dosher, 1994).

Ultimately, the development of kinetic stereoscopic displays is one of the long term goals of this research. Effectively disambiguating the depth plane in KDEX displays can lead to obvious benefits in terms of the future reliability and usability of the overall system.

Finally, as has already been mentioned, it would be beneficial if future experiments could be conducted with the participation of real life airport security personnel. Consequently, the difficulty level of the bags could potentially be significantly increased. For example, handguns could potentially be disassembled and their components hidden in different locations inside each bag, or even across several bags. Also, a broader variety of targets could be used. An example would be clandestine improvised explosive devices (IED), or maybe even drug and other contraband concealments. This would be permissible as professional screeners would have already been trained how to recognise them. Additionally, utilising expert screeners would remove the burden of having to train novice participants on the basics of x-ray security screening.

Additionally, as part of conducting research with professional screeners, an evaluation of the relevant considerations, arising from the use of kinetic stimuli in the actual security task, could be undertaken. This would facilitate making these potential experiments as realistic as possible. For instance in the future a study could perhaps be conducted in the field (e.g. at an airport); this would complement the current laboratory based research. In this context additional real task parameters such as maintaining vigilance and alertness in the face of the low occurrence of real targets, while looking at kinetic x–ray images, can also be investigated.

One other possible avenue of research is to investigate if there are any long term effects on screeners from viewing kinetic stimuli, as part of their normal working routine. Also, as mentioned earlier, environmental factors, such as background noise, could be looked into, as these are not parameters in a laboratory setting. Although these issues were beyond the scope of the current research it would be beneficial if they were to be investigated as part of future field research.

4.6 Closing Remarks

The experiments described in this thesis represent years of intensive interdisciplinary work by the author, in collaboration with the University team. The x-ray scanner had to be refurbished, stimuli collected, and experiment software had to be developed. Furthermore, several threatening objects, as well as numerous travel items and various pieces hand luggage, were sourced and procured, and a multitude of KDEX sequence scans were acquired. Additionally, almost 100 participants were recruited and trained by the author in order to complete these experiments. In total they completed close to 58,000 valid experimental trials; this represented hundreds of man-hours of research work.

In conclusion, it is felt that this line of research adequately fulfilled the goals it set out for itself. The research questions were thoroughly evaluated and the data clearly supports the alternative hypothesis. This means that KDEX indeed offers a significant improvement in performance on visual inspection tasks, over existing technology. Furthermore, this technique is the most ergonomic way of presenting luggage scans taken from multiple different angles. The intention is to continue pursuing this research path, with the aim to continue improving the KDEX technique until this system is ready to be deployed in the field.

References

Abidi, B., Zheng, Y., Gribok, A., & Abidi, M. (2005). Screener evaluation of pseudo-colored single energy x-ray luggage images. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Workshop CDROM, San Diego, CA, USA, June 2005.

Anderson, K. C., & Sieger, R. M. (2005). Three-dimensional structure-from-motion selectivity in the anterior superior temporal polysensory area, STPa, of the behaving monkey. *Cerebral Cortex, 15*, 1299-1307.

Arditi, A. (2005). Improving the design of the letter contrast sensitivity test. *Investigative Ophthalmology and Visual Science*. *46*(6), 2225-2229.

Balakrishnan, J. D. & MacDonald, J. A. (2003). Is the area measure a historical anomaly? *Canadian Journal of Experimental Psychology, 57*(3), 238-256.

Bex, P. J., & Dakin, S. C. (2005). Spatial interference among moving targets. *Vision Research, 45*, 1385-1398.

Bradshaw, M. F., Parton, A. D., & Eagle, R. A. (1998). The interaction of binocular disparity and motion parallax in determining perceived depth and perceived size. *Perception, 27*, 1317-1331.

Bradshaw, M. F., Parton, A. D., & Glennerster, A. (2000). The task-dependent use of binocular disparity and motion parallax information. *Vision Research, 40*, 3725-3734.

Braunstein, M. L., & Todd, J. T. (1990). On the distinction between artifacts and information. *Journal of Experimental Psychology: Human Perception and Performance, 16*(1) 211-216.

Caudek, C., & Domini, F. (1998). Perceived orientation of axis of rotation in structure-from-motion. *Journal of Experimental Psychology: Human Perception and Performance, 24*(2), 609-621.

Caudek, C., & Rubin, N. (2001). Segmentation in structure from motion: Modeling and psychophysics. *Vision Research, 41*, 2715-2732.

Chun, M. M., & Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36*, 28-71.

Chung, S., Levi, D., & Legge, G. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research, 41*, 1833-1850.

CNN (2002, March 26). Knives, guns, fake bombs elude airport security. <http://edition.cnn.com/2002/US/03/25/airport.security/>. [Accessed 6 December 2008].

CNN (2003, March 11). TSA: More than 4.8 million items seized. <http://edition.cnn.com/2003/TRAVEL/03/10/airport.security.ap/index.html>. [Accessed 6 December 2008].

Corwin, J. (1994). On measuring discrimination and response bias: Unequal numbers of targets and distractors and two classes of distractors. *Neuropsychology, 8*(1), 110-117.

Criminal Occurrence description. The Aviation Safety Network Website. Last updated: 19 November 2008. <http://aviation-safety.net/database/record.php?id=19941210-0>. [Accessed 19 November 2008].

Day, R. H., & Power, R. P. (1965). Apparent reversal (oscillation) of rotary motion in depth: An investigation and a general theory. *Psychological Review, 72*(2), 117-127.

DeCarlo, L. T. (2003). Source monitoring and multivariate signal detection theory, with a model for selection. *Journal of Mathematical Psychology, 47*, 292–303.

Dosher, B. A., Landy, M. S., & Sperling, G. (1989). Ratings of kinetic depth in multidot displays. *Journal of Experimental Psychology: Human Perception and Performance, 15*(4), 816-825.

Dougherty, B. E., Flom, R. E. & Bullimore, M. A. (2005). An evaluation of the mars letter contrast sensitivity test. *Optometry and Vision Science, 82*(11), 970-975.

Durgin, F. H., Proffitt, D. R., Olson, T. J., & Reinke, K. S. (1995). Comparing depth from motion with depth from binocular disparity. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 679-699.

Eckstein, M. P., Ahumada, A. J. Jr., & Watson, A. B. (1997). Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise. *Journal of the Optical Society of America, 14*(9), 2406-2419.

Evans, J. P. O. (2002). Stereoscopic imaging using folded linear dual-energy x-ray detectors. *IOP Journal of Measurement Science & Technology, 13*, 1388-1397.

Evans, J. P. O. (2003). Kinetic depth effect X-ray (KDEX) imaging for security screening. *International Conference on Visual Information Engineering (VIE 2003)*, July 2003, 69- 72.

Evans, J. P. O., & Hon, H. W. (2002). Dynamic stereoscopic X-ray imaging. *Journal of Non-destructive Testing and Evaluation, 35*, 337-345.

Fang, F., & He, S. (2004). Stabilized structure from motion without disparity induces disparity adaptation. *Current Biology, 14*, 247-251.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*, 861–874.

Fernandez, J. M., & Farell, B. (2006). A reversed structure-from-motion effect for simultaneously viewed stereo surfaces. *Vision Research, 46*, 1230-1241.

Fleck, M. S., & Mitroff, S. R. (2007). Rare targets are rarely missed in correctable search. *Psychological Science, 18*(11), 943-947.

Fleck, M., & Mitroff, S. (2008). Videogamers excel at finding rare targets [Abstract]. *Journal of Vision*, *8*(6), 313, 313a.

Fiore, S. M., Jentsch, F., & Rubinstein, J. (2004). Human-centered research for keeping our nation safe. *Cognitive Technology, 9*(2), 20-21.

Fiore, S. M., Rubinstein, J., & Jentsch, F. (2004). Considering science and security from a broader research perspective. *Cognitive Technology, 9*(2), 40-42.

Fiore, S. M., Scielzo, S., & Jentsch, F. (2004). Stimulus competition during perceptual learning: Training and aptitude considerations in the x-ray security screening process. *Cognitive Technology, 9*(2), 34-39.

Furmanski, C. S., & Engel, S. A. (2000). Perceptual learning in object recognition: Object specificity and size invariance. *Vision Research, 40*, 473-484.

Gale A. G., Mugglestone, M. D, Purdy, K..J., & McClumpha, A. (2000). Is airport baggage inspection just another medical image? In Krupinski E.A. (Ed.) *Medical Imaging: Image Perception and Performance. Progress in Biomedical Optics and Imaging, 1*(26), 184-192.

Gale, A.G., Purdy, K. & Wooding, D. (2005). *Designing out terrorism: Human factors issues in airport baggage inspection*. In: Human Factors Design, Safety, and Management. Maastricht: Shaker.

Gale, T. M., Done, D. J., & Frank, R. J. (2001). Visual crowding and category specific deficits for pictorial stimuli: A neural network model. *Cognitive Neuropsychology, 18*(6), 509-550.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96*(2), 267-314.

Geisler, W. S., Perry, J. S., & Najemnik, J. (2006). Visual search: The role of peripheral information measured using gaze-contingent displays. *Journal of Vision, 6*, 858-873.

Gibson, J. J. (1950). *The perception of the visual world.* Boston: Houghton Mifflin Co.

Gibson, J. J., & Gibson, E. J. (1957). Continuous perspective transformations and the perception of rigid motion. *Journal of Experimental Psychology, 54*(2), 129-138.

Green, B. F. (1961). Figure coherence in the kinetic depth effect. *Journal of Experimental Psychology, 62*(3), 272-282.

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*(1), 88-94.

Green, S., Blumenstein, M., Muthukkumarasamy, V., & Jo, J. (2005). Investigation of a classification-based technique to detect illicit objects for aviation security. *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, Innsbruck, Austria, February 2005.

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley. Reprinted (1974). Huntington, NY: Kreiger.

Greenwald, A. G., Oakes, M. A. & Hoffman, H. G. (2003). Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology, 39*, 399-405.

Gogel, W. C. (1964). Perception of depth from binocular disparity. *Journal of Experimental Psychology, 67*(4), 379-386.

Haijiang, Q., Saunders, J. A., Stone, R. W., & Backus, B. T. (2006). Demonstration of cue recruitment: Change in visual appearance by means of Pavlovian conditioning. *PNAS, 103*(2), 483-488.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Harris, D. H. (2002). How to really improve airport security. *Ergonomics in Design*, 17-22.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload*, 139-183. Amsterdam: Elsevier.

Harvey, L. O., Jr. (1992). The critical operating characteristic and the evaluation of expert judgment. *Organizational Behavior & Human Decision Processes, 53*(2), 229–251.

Harvey, L. O. Jr., Hammond, K. R., Lusk, C. M., & Mross, E. F. (1992). The Application of signal detection theory to weather forecasting behaviour. *Monthly Weather Review, 120*, 863-883.

Haymes, S. A., Roberts, K. F., Cruess, A. F., Nicolela, M. T., LeBlanc, R. P., Ramsey, M. S., Chauhan, B. C., & Artes, P. H. (2006). The letter contrast sensitivity test: Clinical evaluation of a new design. *Investigative Opthalmology and Visual Science, 47*(6), 2739-2745.

Hayward, W. G. (1998). Effects of outline shape in object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 24*(2), 427-440.

Helmholtz, H. von. (1925). *Treatise on Physiological Optics,* vol. 3. Trans. From 3rd German ed. J. P. C. Southall, ed. Opt. Soc. Amer. Republished (1962). New York: Dover.

Herrmann, D. (1996). Cognitive technology's contribution to cognitive psychology: The optimization of discovery and application. *Cognitive Technology, 1*(1), 4-6.

Hershberger, W. A., Stewart, M. R., & Laughlin, N. K. (1976). Conflicting motion perspective simulating simultaneous clockwise and counterclockwise rotation in depth. *Journal of Experimental Psychology: Human Perception and Performance, 2*(2), 174-178.

Hilscher, M. B. (2005). *Performance implications of alternative color-codes in airport security screening*. Unpublished doctoral dissertation, University of Central Florida, Orlando, USA.

Hochberg, J., & Peterson, M. A. (1987). Piecemeal organization and cognitive components in object perception: Perceptually coupled responses to moving objects. *Journal of Experimental Psychology: General, 116*(4), 370-380.

Hon, H. W., & Evans, J. P. O. (2002). Multiple view line-scan imaging. *IEE Proceedings on Optoelectronics*, *149*(2), 45-50.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual search has no memory. *Nature, 394*, 575-577.

Howard, I. P. (2002). *Seeing in depth, Vol. 1: Basic mechanisms*. Toronto, Canada: I Porteus.

Hubona, G. S., Wheeler, P. N., Shirah, G. W., & Brandt, M. (1999). The role of object shadows in promoting 3D visualization. *ACM Transactions on Computer-Human Interaction, 6*, 1-21.

Huckauf, A., & Heller, D. (2004). On the relations between crowding and visual masking. *Perception and Psychophysics, 66*(4), 584-595.

Ito, H. (1999). Two processes in stereoscopic apparent motion. *Vision Research, 39*, 2739-2748.

Ishihara, S. (1917). *Tests for colour-blindness*. Handaya, Tokyo: Hongo Harukicho.

Jusczyk, P. W., Johnson, S. P., Spelke, E. S. & Kennedy, L. J. (1999). Synchronous change and perception of object unity: Evidence from adults and infants. *Cognition, 71*, 257-288.

Kase, K. (2002). *Effective use of color in x-ray image enhancement for luggage inspection*. Unpublished master's thesis, The University of Tennessee, Knoxville, USA.

Katkov, M., Tsodyks, M., & Sagi, D. (2006). Singularities in the inverse modeling of 2AFC contrast discrimination data. *Vision Research, 46*, 259-266.

Kaufman, L. (1974). *Sight and Mind*. New York, USA: Oxford University Press.

Kim, C. Y., & Blake, R. (2005). Psychophysical magic: Rendering the visible 'invisible'. *Trends in Cognitive Science, 9*(8), 381-388.

Krupinski, E. A., Berger, W. G., Dallas, W. J., & Roehrig, H. (2003). Searching for nodules: what features attract attention and influence detection? *Academic Radiology, 10*, 861-868.

Krupinski, E. A., & Roehrig, H., (2002). Pulmonary nodule detection and visual search: P45 and P104 monochrome versus color monitor displays. *Acaemicd Radiology, 9*(6), 638-45.

Kunar, M. A., Flusberg, S. J., & Wolfe, J. M. (2006). Contextual cueing by global features. *Perception & Psychophysics, 68*, 1204-1216.

Leone, K., & Liu, R. (2005). The key design parameters of checked baggage security screening systems in airports. *Journal of Air Transport Management, 11*, 69–78.

Liu, X., & Gale, A. G. (2007). Effects of time pressure on searching for terrorist threats in x-ray air passenger luggage images. In: *Human Factors Issues in Complex System Performance*, D. de Waard et al. (Eds.), Shaker Publishing, Aachen.

Liu, X., Gale, A.G., Purdy, K., & Song, T. (2006). Is that a gun? The influence and features of bags and threat items on detection performance. In: *Contemporary Ergonomics, Proceedings of the 2006 Ergonomics Society Conference*, 17-22.

Liu, X., Gale, A. G., & Song, T. (2007). Detection of terrorist threats in air passenger luggage: Expertise development. In: *Proceedings of the 41st Annual IEEE International Carnahan Conference on Security Technology*, Ottawa, Canada, 301-306.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's guide*. (2nd ed.). New York: Cambridge University Press.

Manning, D. J. (1998). Evaluation of diagnostic performance in radiography. *Radiography, 4*, 49-60.

Manning, D. J., Gale, A., & Krupinski, E. A. (2005). Perception research in medical imaging. *The British Journal of Radiology, 78*, 683–685.

Marzban, C. (2004). The ROC curve and the area under it as a performance measure. *Weather and Forecasting, 19*(6), 1106-1114.

McCarley, J. S., & Carruth, D. W. (2004). Occulomotor scanning and target recognition in luggage x-ray screening. *Cognitive Technology, 9*(2), 26-29.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science, 15*(5), 302-306.

McCormick, E. J., & Sanders, M. S. (1982). *Human Factors in Engineering and Design*. (5th ed.). New York: McGraw-Hill.

Meneer, T., Phillips, L., Donelly, N., Barrett, D. J. K., & Cave, K. R. (2004). Search efficiency for multiple targets. *Cognitive Technology, 9*(2), 22-25.

Metz, C. E. (2006). Receiver operating characteristic (ROC) analysis: A tool for quantitative evaluation of observer performance and imaging systems. *Journal of the American College of Radiology, 3*, 413-422.

Metzger, W. (1934). Tiefenerscheinungen in optischen Bewegungsfeldern. *Psychologische Forschung, 20*, 195-260.

Michel, S., Koller, S. M., de Ruiter, J. C., Moerland, R., Hogervorst, M., & Schwaninger, A. (2007). Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners. *41st Annual IEEE International Carnahan Conference on Security Technology*, 201-206.

Miles, W. R. (1931). Movement interpretation of the silhouette of a revolving fan. *American Journal of Psychology, 43*, 392-405.

Mukai, I., & Watanabe, T. (1999). The influence of structure from motion on motion correspondence. *Perception, 28*, 331-340.

Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on x-ray image processing. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*(3), 553-557.

NASA -Task Load Index (TLX): *Computerized Version.* (1986). Moffett Field. CA: NASA-Ames Research Center, Aerospace Human Factors Research Division.

Nawrot, M. (2003). Depth from motion parallax scales with eye movement gain. *Journal of Vision, 3*, 841-851.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387-391.

Necker, L. A. (1832). Observations on some remarkable optical phaenomena seen in Switzerland; and on an optical phaenomenon which occurs on viewing a figure of a crystal or geometrical solid. *London and Edinburgh Philosophical Magazine and Journal of Science*, *1*(5), 329–337.

Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology, 229*, 3–8.

Oertel, C., & Bock, P. (2006). Identification of objects-of-interest in x-ray images. *Proceedings of the 35th Applied Imagery and Pattern Recognition Workshop*, October 2006.

Olzak, L. A., & Thomas, J. P. (1986). *Seeing spatial patterns*. In K. R. Boff, L. Kaufmann, & J. P. Thomas (Eds.), *Handbook of human perception and performance: Vol. 1. Sensory processes and perception*. New York: Wiley.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*, 1227-1268.

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin and Review, 1*(1), 29-55.

Paradis, A. L., Cornilleau-Peres, V., Droulez, J., Van de Moortele, P. F., Lobel, E., Berthoz, A., Le Bihan, D., & Poline, J. B. (2000). Visual perception of motion and 3-D structure from motion: An fMRI study. *Cerebral Cortex, 10*, 772-783.

Park, S. H., Goo, J. M., & Jo, C. H. (2004). Receiver operating characteristic (ROC) curve: Practical review for radiologists. *Korean Journal of Radiology, 5*, 11-18.

Park, H., Quinlan, J., Thornton, E., & Reder, L. M. (2004). The effect of midazolam on visual search: Implications for understanding amnesia. *PNAS, 101*(51), 17879-17883.

Parker, K. J., Zhang, M., & Rubens, D. J. (2006). An introduction to color in medical imaging. *Journal of Imaging Science and Technology, 50*(1), 12-16.

Pelli, D. G., Cavanagh, P., Desimone, R., Tjan, B., & Treisman, A. (2007). Crowding: Including illusory conjunctions, surround suppression, and attention. *Journal of Vision*, *7*(2): i, 1.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*, 1136-1169.

Pelli, D. G., & Tillman. K. A. (2008). The uncrowded window of object recognition. *Nature neuroscience, 11*(10), 1129-1135.

Peterson, M. S., & Kramer, A. F. (2001). Attentional guidance of the eyes by contextual information and abrupt onsets. *Perception and Psychophysics, 63*(7), 1239-1249.

Peterson, M. A., & Gibson, B. S. (1994). Must figure-ground organization precede object recognition? An assumption in peril. *Psychological Science, 5*(5), 253-259.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research, 41*, 3145-3161.

Prism 4 for Windows (Version 4.03). (2005). [Computer software]. San Diego, CA: Graph Pad Software.

Proffitt, D. R., Hecht, H., Rock, I., & Schubert, J. (1992). Stereokinetic effect and its relation to the kinetic depth effect. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 3–21.

Ratcliff, R. & McKoon, G. (1996). Bias effects in implicit memory tasks. *Journal of Experimental Psychology: General, 125*(4), 403–421.

Ringach, D. L., Hawken, M. J., & Shapley, R. (1995). Binocular eye movements caused by the perception of three-dimensional structure from motion. *Vision Research, 36*(10), 1479-1492.

Rock, I. (1975). *An Introduction to Perception*. New York: Macmillan.

Rock, I., & Smith, D. (1981). Alternative solutions to kinetic stimulus transformations. *Journal of Experimental Psychology: Human Perception and Performance, 7*(1), 19-29.

Rogers, B. J., & Collett, T. S. (1989). The appearance of surfaces specified by motion parallax and binocular disparity. *The Quarterly Journal of Experimental Psychology, 41A*(4), 697-717.

Roling, P., Sowden, P., Davies, I., Ozgen, E., & Lawler, M. (1997). Perceptual learning in inspection tasks. *Engineering Psychology and Cognitive Ergonomics, 2*, 121-128.

Schmuckler, M. A., & Proffitt, D. R. (1994). Infants' perception of kinetic depth and stereokinetic displays. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 122-130.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners: Visual abilities and visual knowledge measurement. *IEEE A&E Systems Magazine, 20*(6), 29-35.

Schwartz, B. J., & Sperling, G. (1983). Luminance controls the perceived 3-D structure of dynamic 2-D displays. *Bulletin of the Psychonomic Society, 21*(6), 456-458.

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory & Cognition, 34*(1), 125-137.

Smeets, M., Ingleby, D., Hoek, H., & Panhuysen, G. (1999). Body size perception in anorexia nervosa: A signal detection theory approach. *Journal of Psychosomatic Research, 46*(5), 465-477.

Smith, D. J., Redford, J. S., Washburn, D. A., & Tagliatela, L. A. (2005). Specific-token effects in screening tasks: Possible implications for aviation security. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(6), 1171-1185.

Sobania, A., & Evans, J. P. O. (2005). Morphological corner detector using paired triangular structuring elements. *Journal of the Pattern Recognition Society*, *38*, (7), 1087-1098.

Sowden, P. T., Davies, I. R. L., Roling, P. (2000). 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*(1), 379-390.

Sperling, G., & Dosher, B. A. (1986). Strategy and optimization in human information processing. In K. R. Boff, L. Kaufmann, & J. P. Thomas (Eds.), *Handbook of human perception and performance: Vol. 1. Sensory processes and perception*. New York: Wiley.

Sperling, G., & Dosher, B. A. (1994). Depth from motion. In *Early Vision and Beyond* (pp. 133-142). Cambridge, MA: MIT Press.

Sperling, G., Dosher, B. A., & Landy, M. S. (1990). How to study the kinetic depth effect experimentally. *Journal of Experimental Psychology: Human Perception and Performance, 16*(2), 445-450.

Sperling, G., Landy, M. S., Dosher, B. A., & Perkins, M. E. (1989). Kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance, 15*(4), 826-840.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401-409.

Thayaparan, K., Crossland, M. D., & Rubin, G. S. (2007). Clinical assessment of two novel contrast sensitivity charts. *British Journal of Ophthalmolology, 91*, 749-752.

Thompson, A. M., & Nawrot, M. (1999). Abnormal depth perception from motion parallax in amblyopic observers. *Vision Research, 39*, 1407-1413.

Tauber, E. S., & Kaufman, L. (1977). Fixation and the stereokinetic phenomenon. *Perception & Psychophysics,* 22, 241-244.

Todd, J. T. (1985). Perception of structure from motion: Is projective correspondence of moving elements a necessary condition? *Journal of Experimental Psychology: Human Perception and Performance, 11*(6), 689-710.

Toet, A., & Franken, E. M. (2003). Perceptual evaluation of different image fusion schemes. *Displays, 24,* 25–37.

Treue, S., Andersen, R. A., Ando, H., & Hildreth, E. C. (1995). Structure-from-motion: Perceptual evidence for surface interpolation. *Vision Research, 35*, 139-148.

Turner, J., Braunstein, M. L., & Andersen, G. J. (1997). Relationship between binocular disparity and motion parallax in surface detection. *Perception & Psychophysics, 59*(3), 370–380.

Tyler, C. W. (2004). Representation of stereoscopic structure in human and monkey cortex. *Trends in Neurosciences, 27*(3), 116-118.

Ullman, S. (1979). The Interpretation of Structure from Motion. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 203*(1153), 405-426.

Vassiliades, V., Evans, P., Kaufman, L., Chan, J.W., & Downes, D. (2008). Preliminary evaluation of kinetic depth x-ray imagery for aviation security screening. *Cognitive Technology*, 13(1), 18-25.

Vlaskamp, B., & Hooge, I. (2006). Crowding degrades saccadic search performance. *Vision Research, 46*, 417-425.

Wagner, R. F., Beam, C. A., & Beiden, S. V. (2004). Reader Variability in mammography and its implications for expected utility over the population of readers and cases. *Medical Decision Making, 24*, 561-572.

Wallach, H., & Karsh, E. B. (1963). The modification of Stereoscopic Depth-Perception and the Kinetic Depth-Effect. *The American Journal of Psychology, 76*(3), 429-435.

Wallach, H., & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology*, *45*(4), 205-217.

Wallach, H., O'Connell, D. N. & Neisser, U. (1953). The memory effect of visual perception of three-dimensional form. *Journal of Experimental Psychology*, *45*(5), 360-368.

Wallach, H., Weisz, A., and Adams, P. A. (1956). Circles and derived figures in rotation. *American Journal of Psychology*, *69*, 48-59.

Washburn, D. A., Taglialatela, L. A., Rice, P. R., & Smith, J. D. (2004). Individual differences in sustained attention and threat detection. *Cognitive Technology, 9*(2), 30-33.

Wickens, T. D. (2002). *Elementary Signal Detection Theory*. New York: Oxford University Press.

Wolfe, J. (1998). Visual search. In Pashler H. (Ed.), *Attention*. London, UK: University College London Press.

Wolfe, J. M., Horowitz, T.S., & Kenner, N. (2005). Rare items often missed in visual searches. *Nature*, *435*, 439-440.

Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S., & Kibbi, N. (2007). Low target prevalence is a stubborn source of errors in visual search tasks. *Journal of Experimental Psychology: General, 136*(4), 623–638.

Appendix A

KDEX Experimental Software

Bespoke experiment software, together with a user-friendly graphical user interface, was developed by the University team for conducting the experimental psychology experiments. The KDEX software code was written in the C++ programming language, and OpenGL was used for graphics handling. The software produced results files in two formats; a text (*.txt) file, and a comma-separated values spreadsheet file (*.csv).

The software evolved to fit the successive changes in the experimental procedure, and thus underwent several revisions. However, the basic software structure can be seen in the two figures below. The tree diagram in Figure A-1 outlines the directory structure of the software. The flowchart in Figure A-2 describes the sequence of core operations in the software.

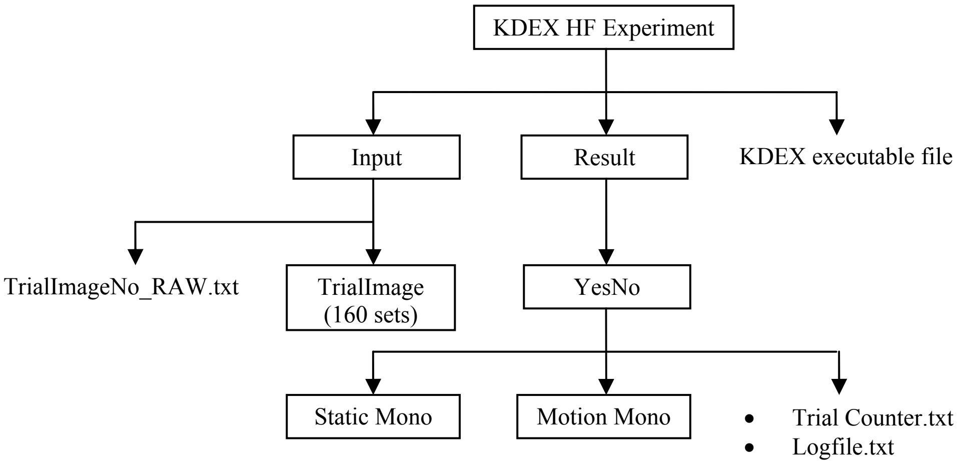


Figure A-1: Directory structure of the KDEX experimental software.

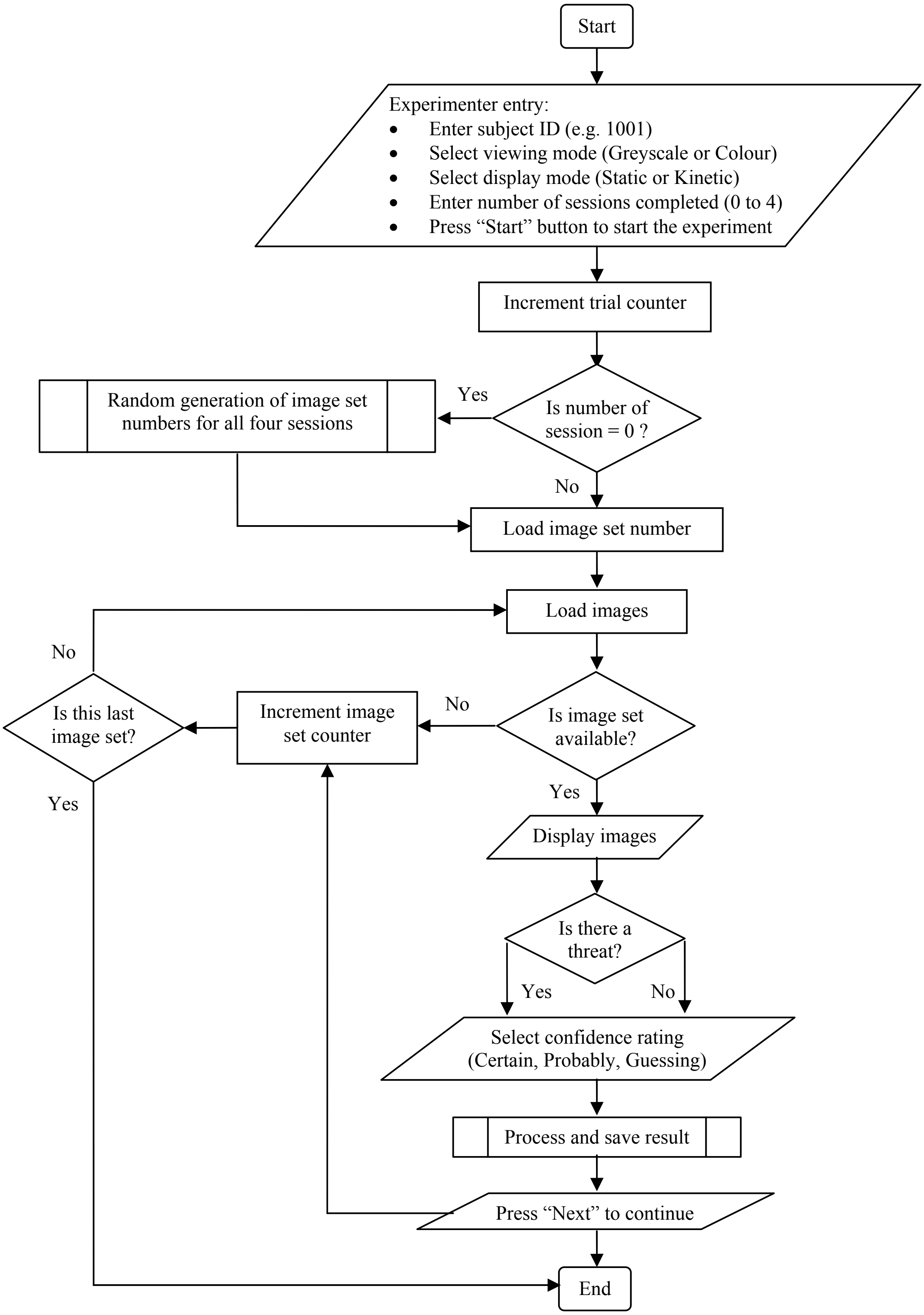

Start
Experimenter entry:
Enter subject ID (e.g. 1001)
Select viewing mode (Greyscale or Colour)
Select display mode (Static or Kinetic)
Enter number of sessions completed (0 to 4)
Press "Start" button to start the experiment
Increment trial counter
Is number of session = 0 ?
Yes
Random generation of image set numbers for all four sessions
No
Load image set number
Load images
No
Is this last image set?
Increment image set counter
No
Is image set available?
Yes
Yes
Display images
Is there a threat?
Yes
No
Select confidence rating (Certain, Probably, Guessing)
Process and save result
Press "Next" to continue
End

Figure A-2. Flowchart of the KDEX experimental software.

Appendix B

Tests Conducted on Study Participants

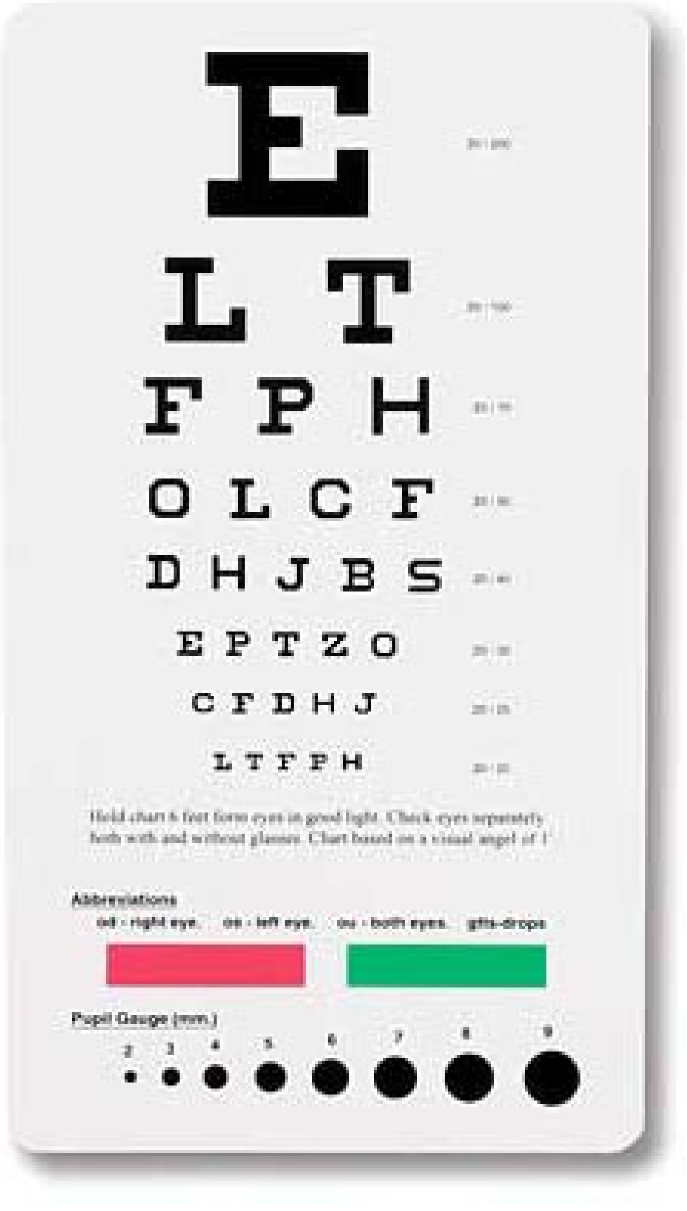

E
20/200
L T
20/100
F P H
20/70
O L C F
20/50
D H J B S
20/40
E P T Z O
20/30
C F D H J
20/25
L T F P H
20/20
Hold chart 6 feet form eyes in good light. Check eyes separately both with and without glasses. Chart based on a visual angel of 1'
Abbreviations
od - right eye. os - left eye. ou - both eyes. gtts-drops
Pupil Gauge (mm.)
2
3
4
5
6
7
8
9

Figure B-1. Mini Snellen pocket eye chart with pupil gauge.

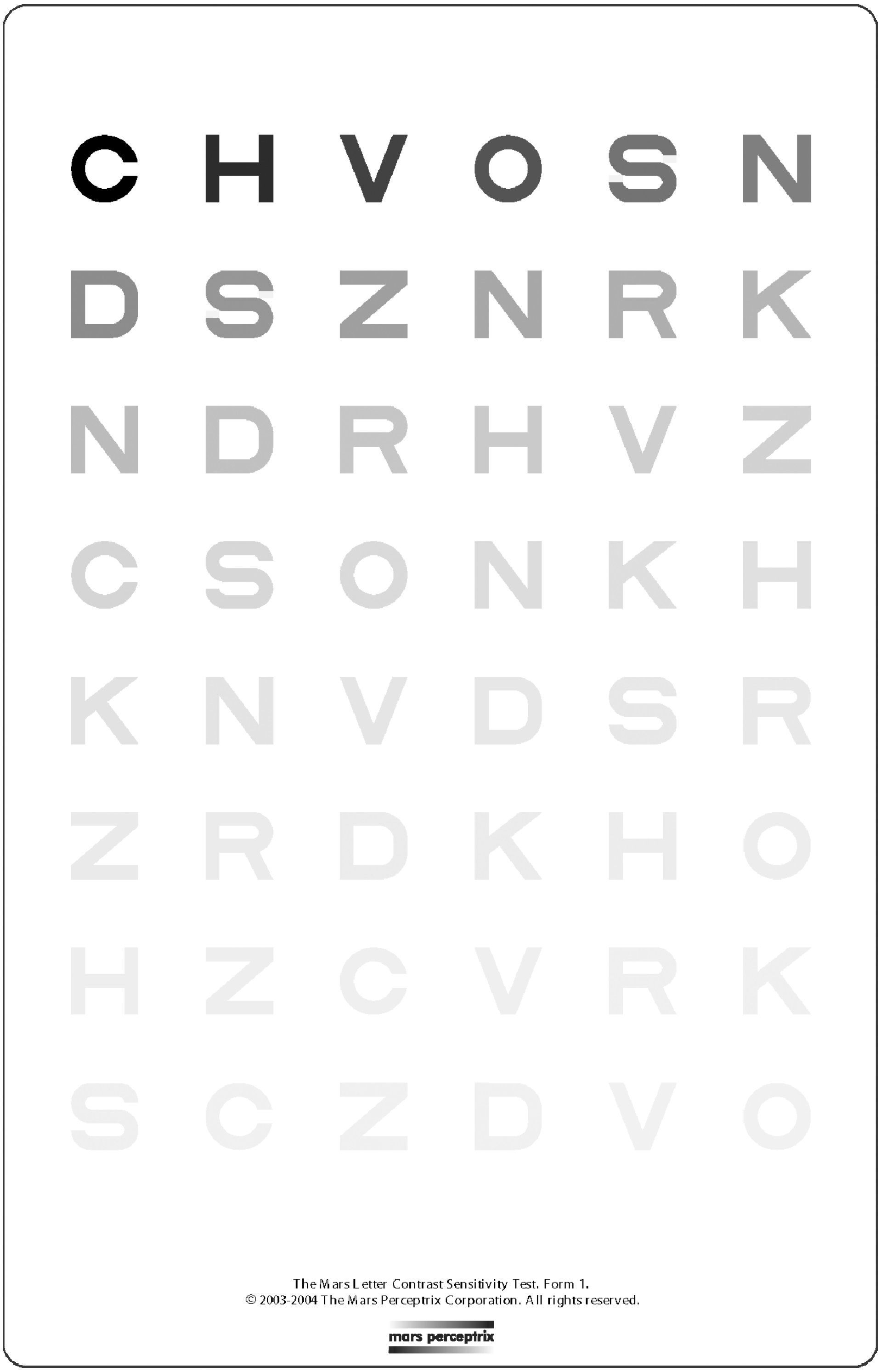

C H V O S N
D S Z N R K
N D R H V Z
C S O N K H
K N V D S R
Z R D K H O
H Z C V R K
S C Z D V O
The Mars Letter Contrast Sensitivity Test. Form 1.
© 2003-2004 The Mars Perceptrix Corporation. All rights reserved.
mars perceptrix

Figure B-2. Mars letter contrast sensitivity test. Here the contrast is enhanced so all the letters are visible. (© 2003-2004 Mars Perceptrix Corporation).

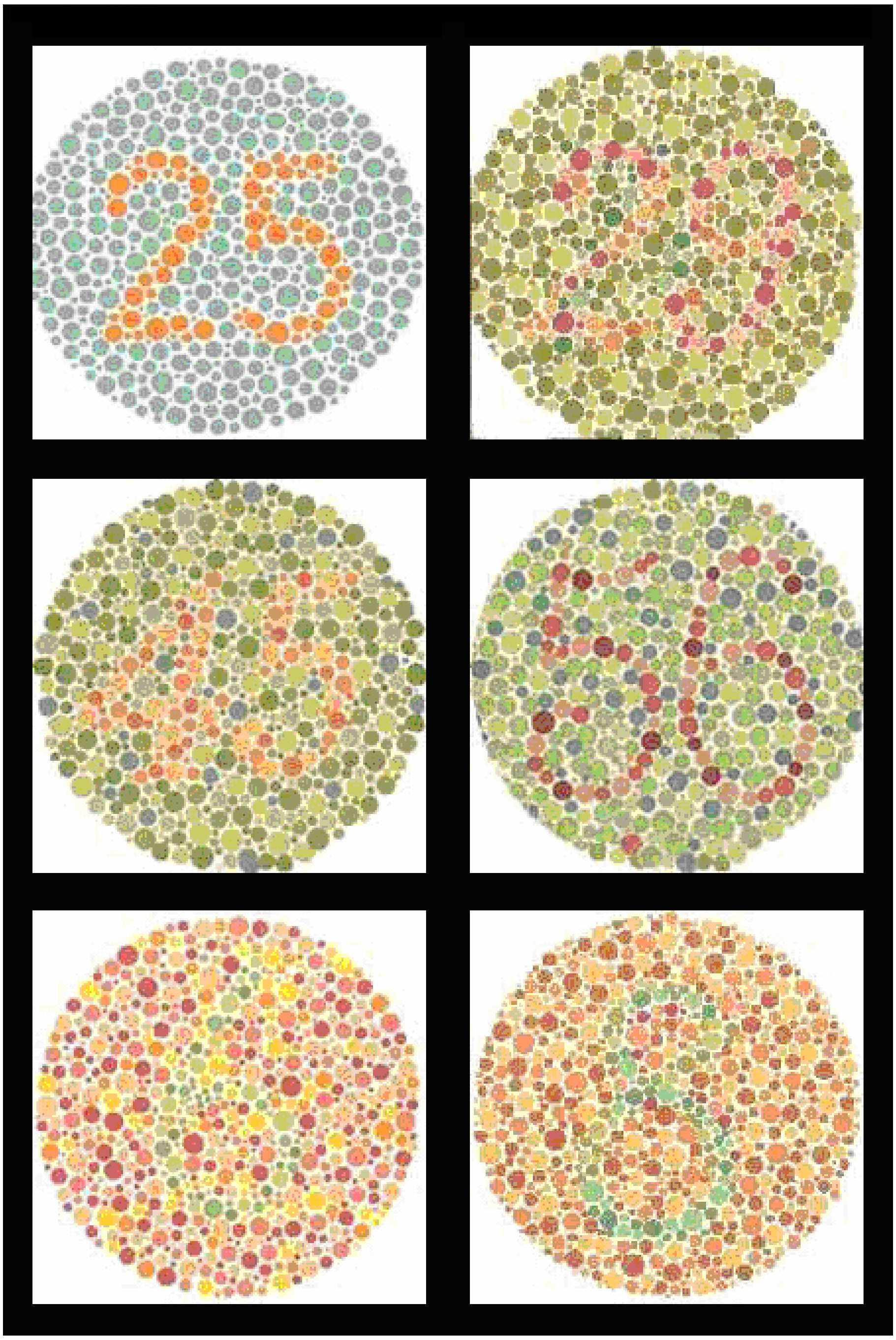
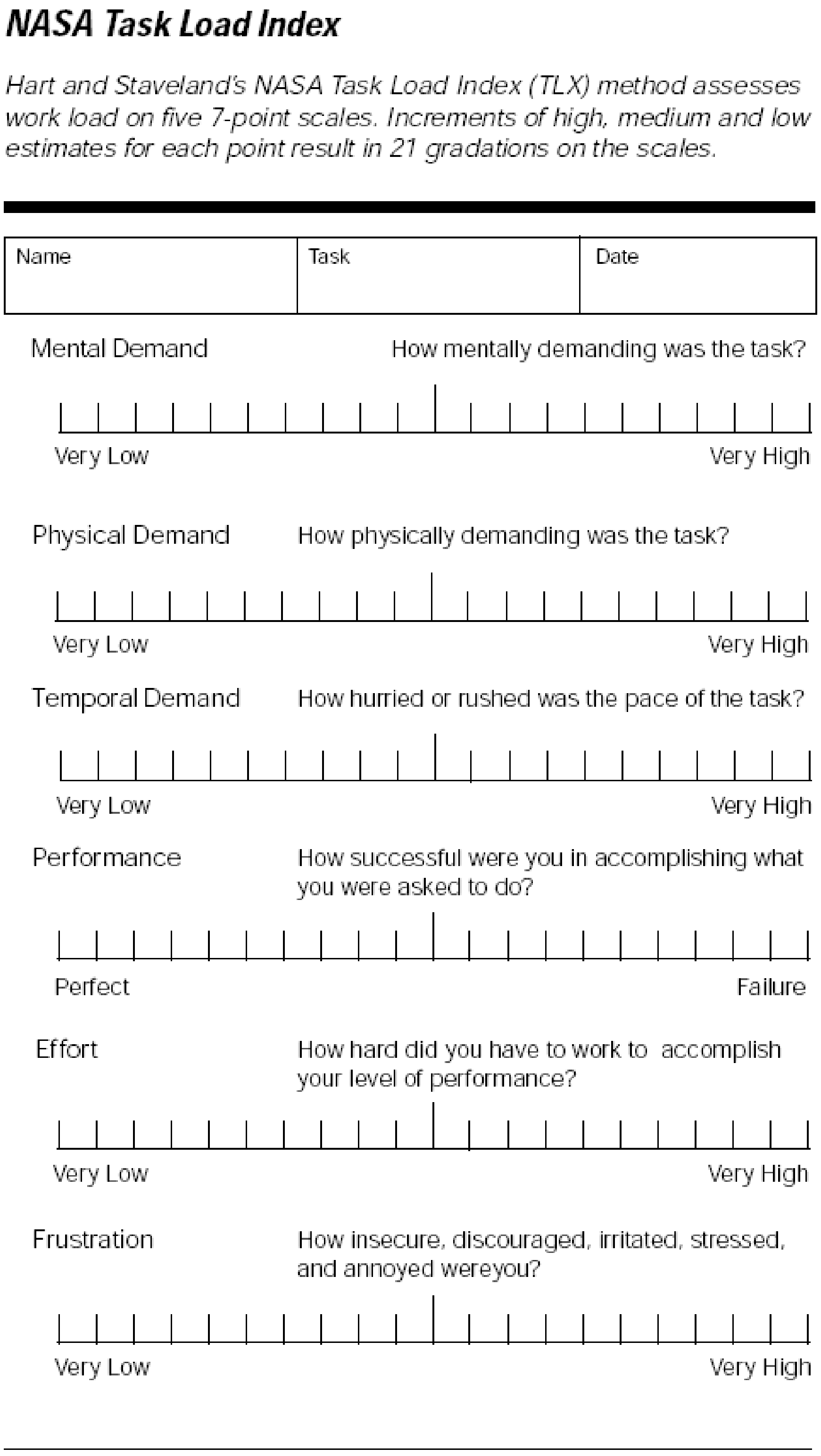


Figure B-3. Ishihara test for colour deficiency. Example of some of the test plates.



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

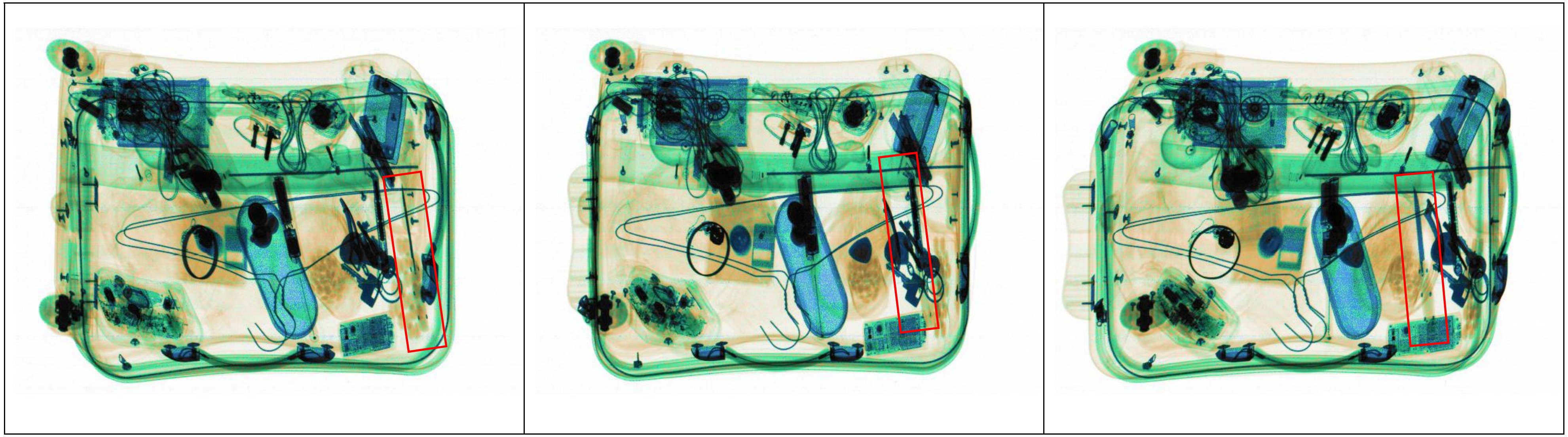
Very Low — Very High

Figure B-4. NASA Task Load Index (TLX). A reliable task workload assessment tool (© 2003 NASA).

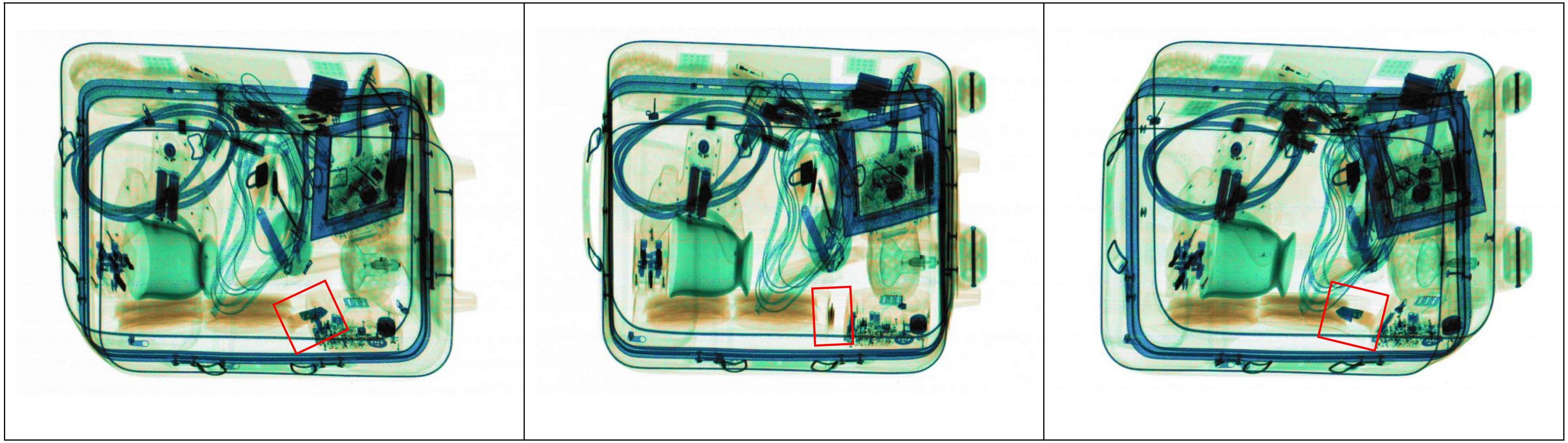
Appendix C

Pseudo-Colour X-Ray Images (Experiments 4 and 5)

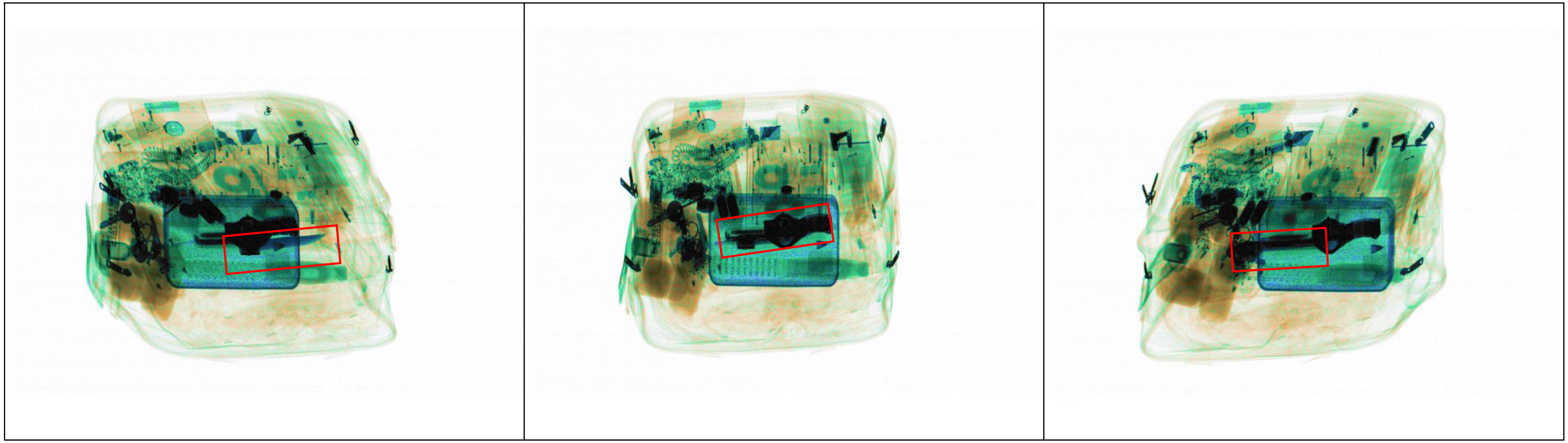
Figure C-1. Experiment Images -20° (Frame 1), 0° (Frame 21), and +20° (Frame 41). The target is marked by the red rectangle.



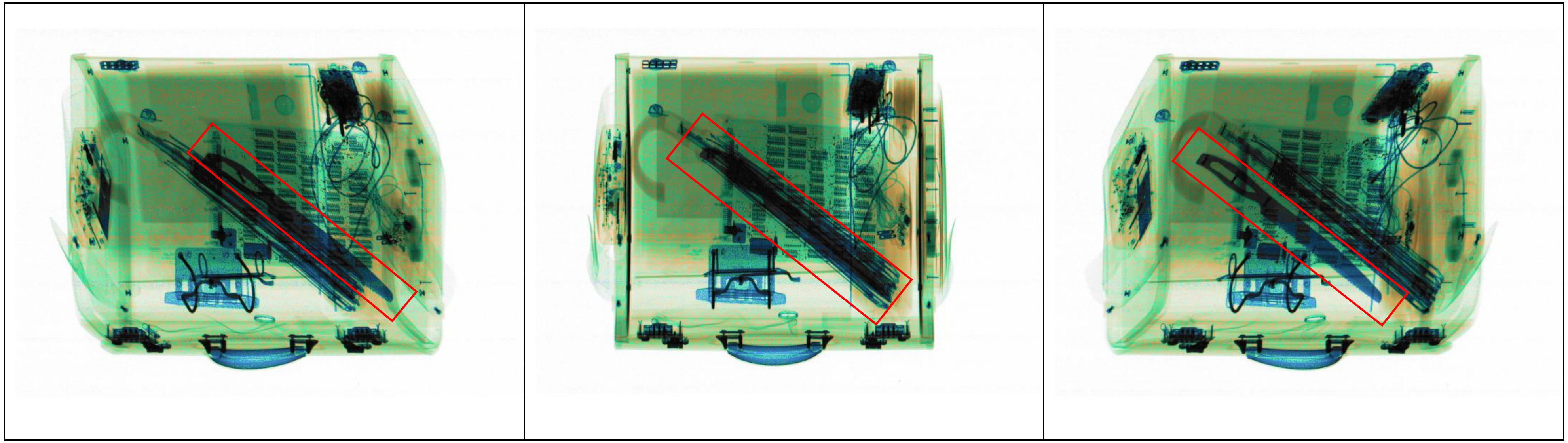
Bag 1: Medium size kitchen knife partially masked, at the periphery of a wheeled bag.



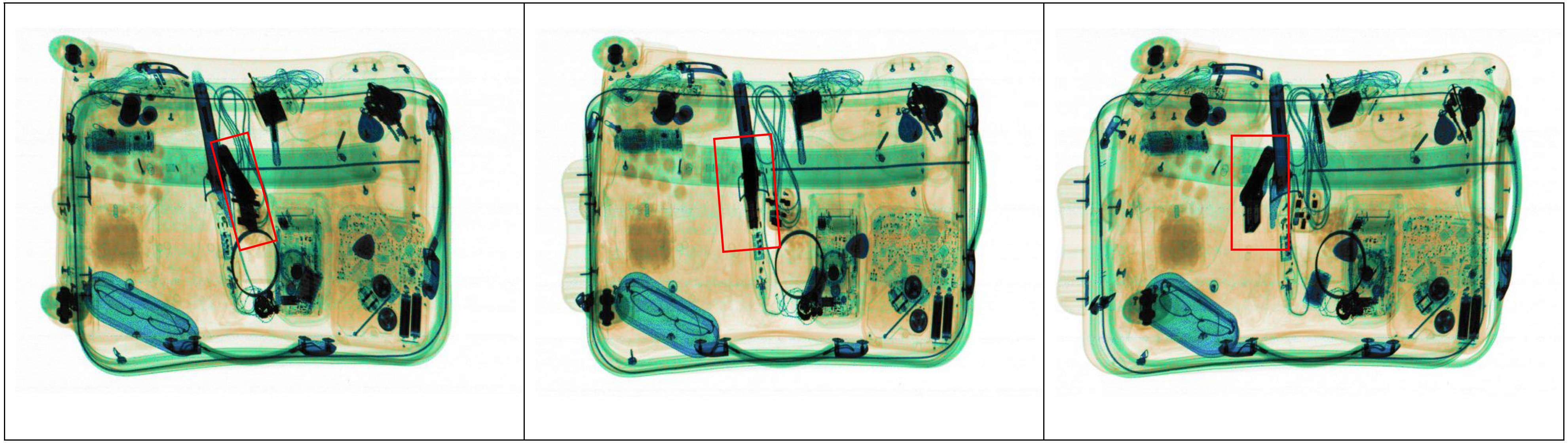
Bag 2: Small size kitchen knife at a minimum footprint with no mask, in the periphery of a wheeled bag.



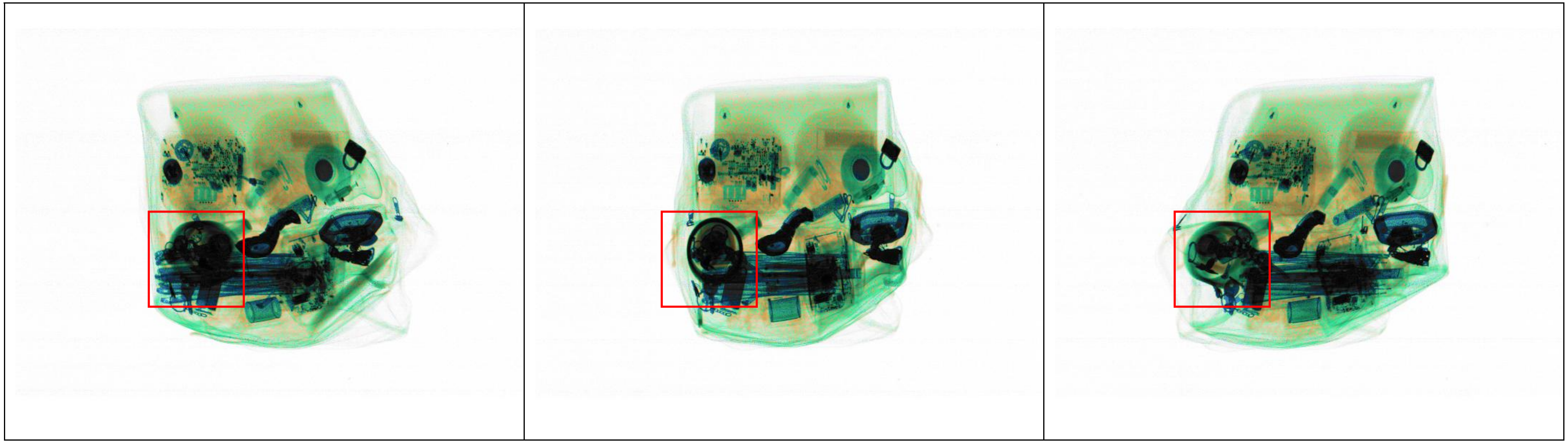
Bag 3: Folding pocket knife partially masked by a watch and pencil case, in the centre of a small soft case bag.



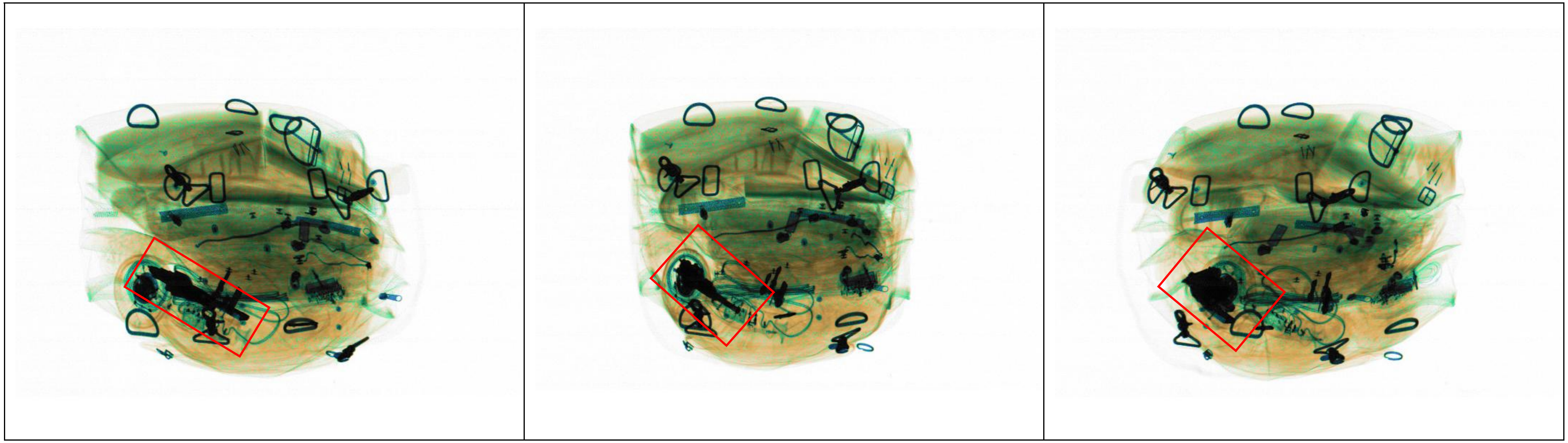
Bag 4: Large bread knife masked by an umbrella, in the centre of a leather briefcase.



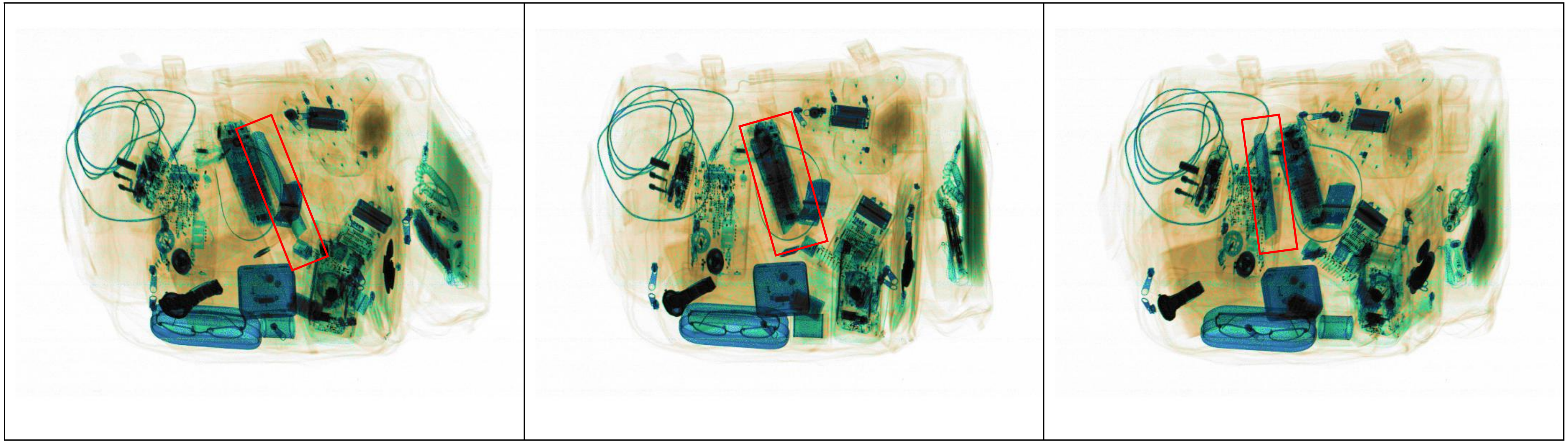
Bag 5: Small size pistol (Baby FN) masked by a curling iron, in the centre of a wheeled bag.



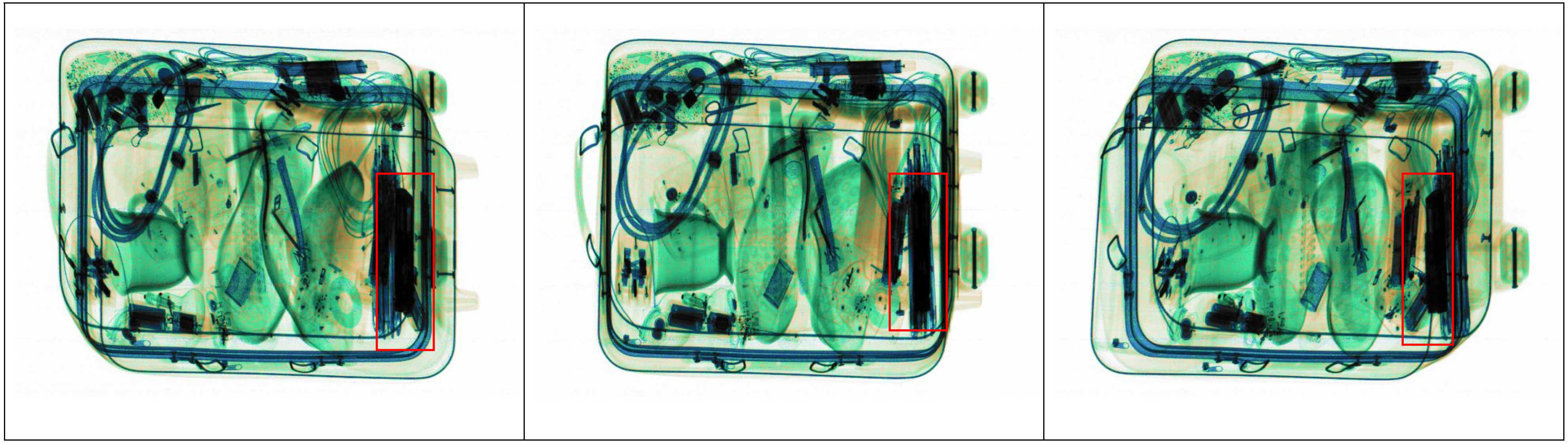
Bag 6: Stun grenade masked by a ceramic coffee mug, at the periphery of a small soft case bag.



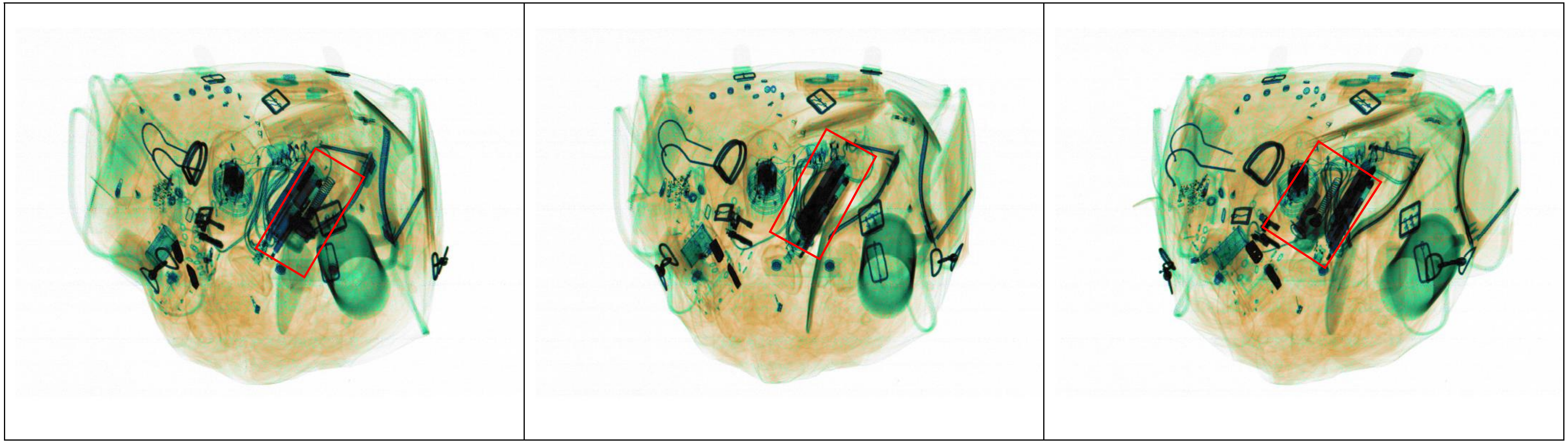
Bag 7: Starting revolver (22 cal) masked by a hair dryer, at the periphery of a canvas bag.



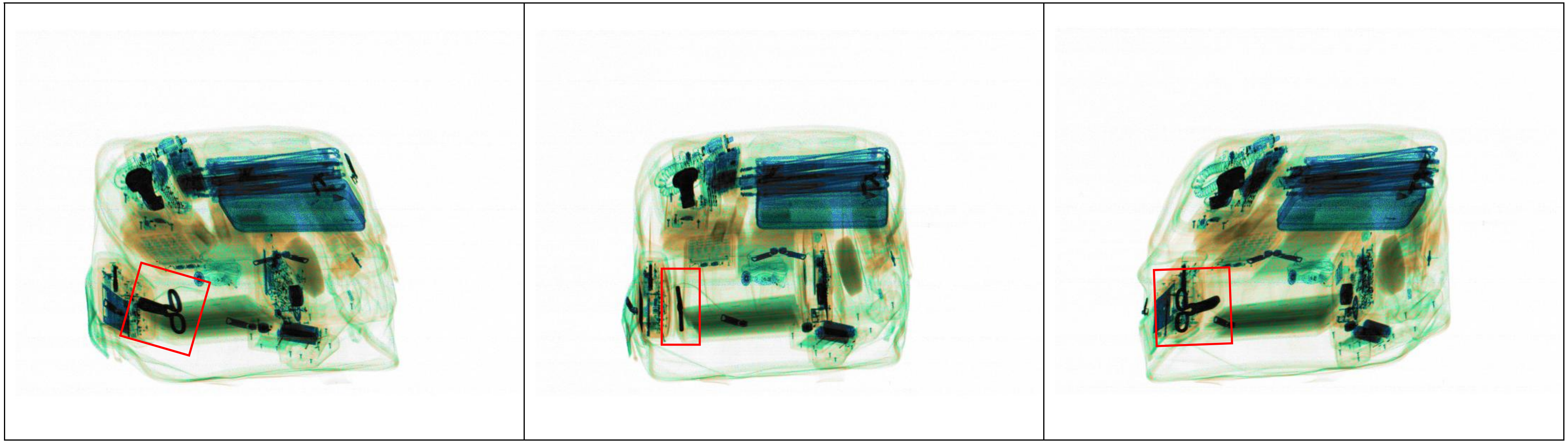
Bag 8: Box cutter (utility knife), masked by a mobile phone, in the centre of a soft case backpack.



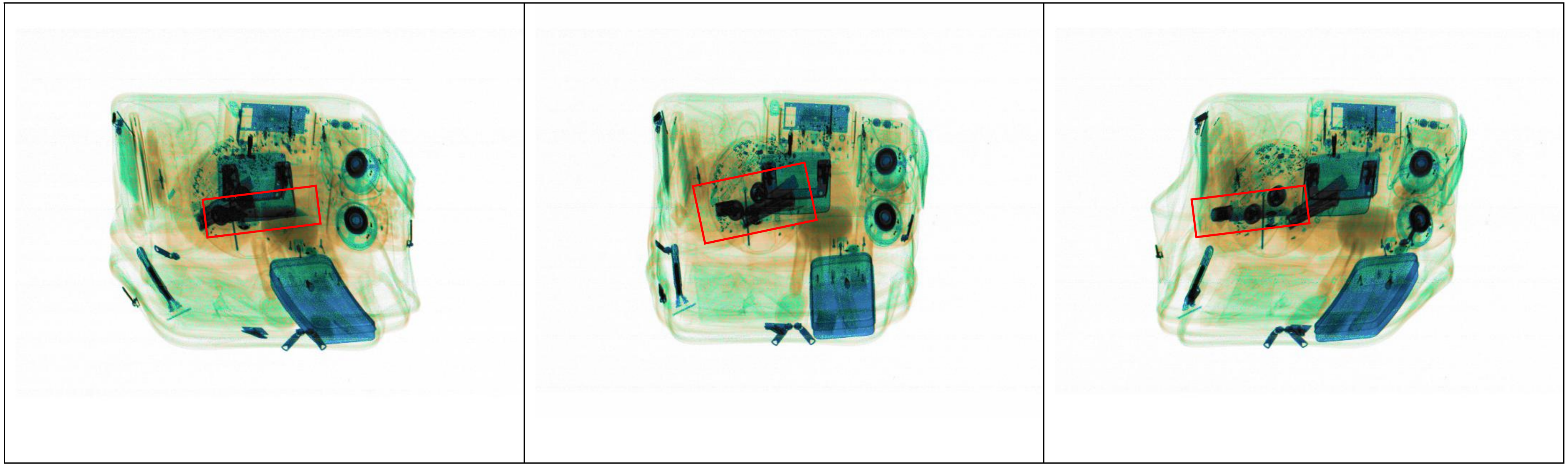
Bag 9: Medium size pistol (Walther P22), completely masked by an umbrella, at the periphery of a wheeled bag.



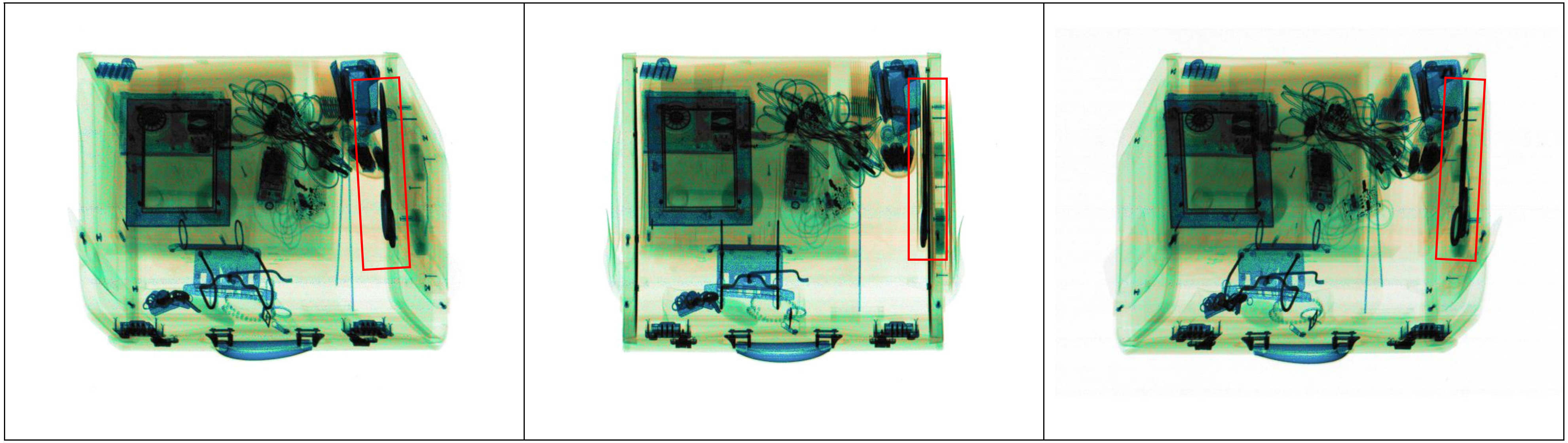
Bag 10: Stun Grenade partially masked by a flashlight, in the centre of a canvas bag.



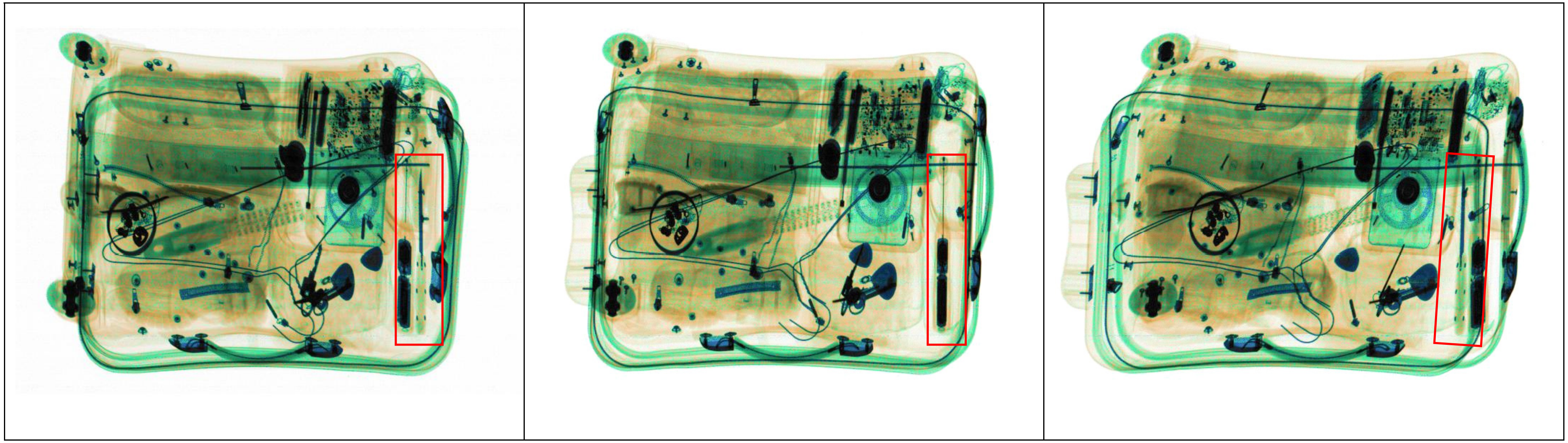
Bag 11: Large metal scissors at a minimum footprint with no mask, at the periphery of a small soft case bag.



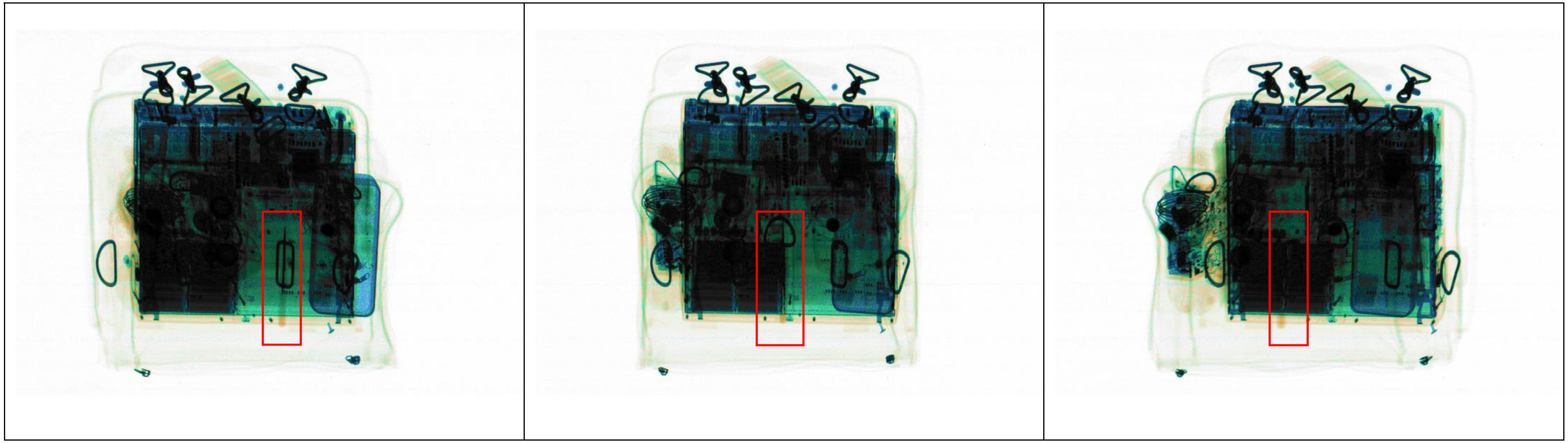
Bag 12: Folding pocket knife partially masked by a stapler, hole-puncher and a CD player, in the centre of a small soft case bag.



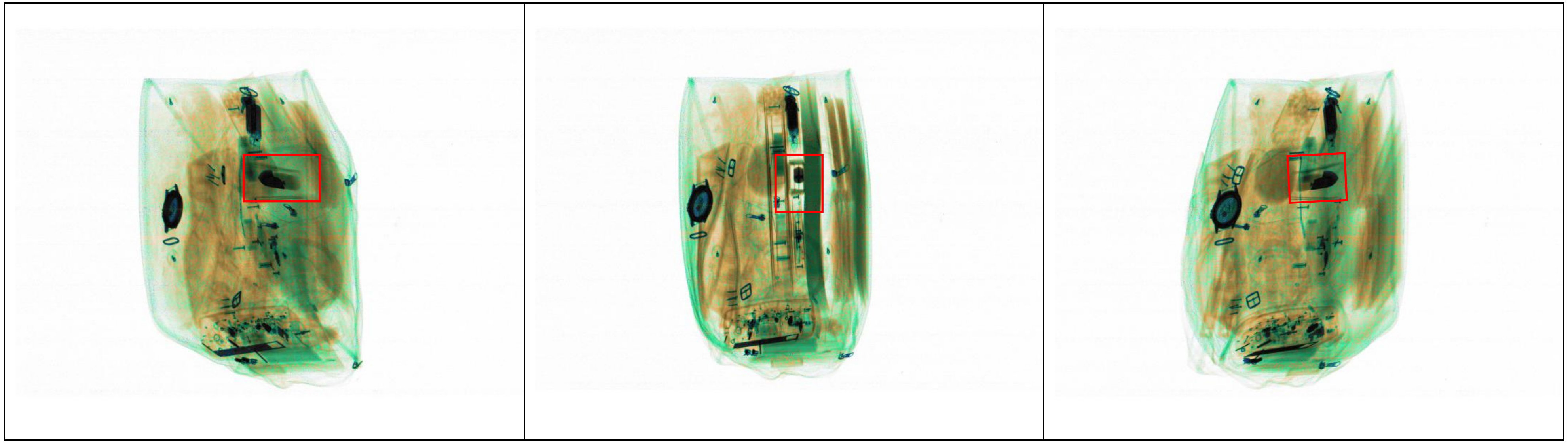
Bag 13: Large metal scissors at a minimum footprint with no mask, at the periphery of a leather briefcase.



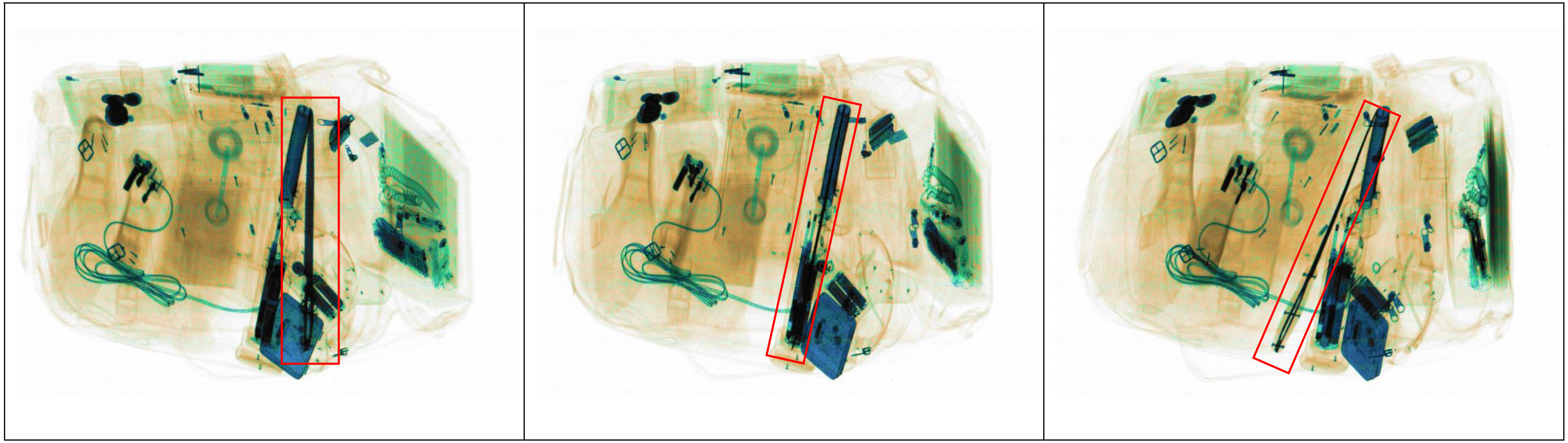
Bag 14: Medium size kitchen knife masked by an electric toothbrush, at the periphery of a wheeled bag.



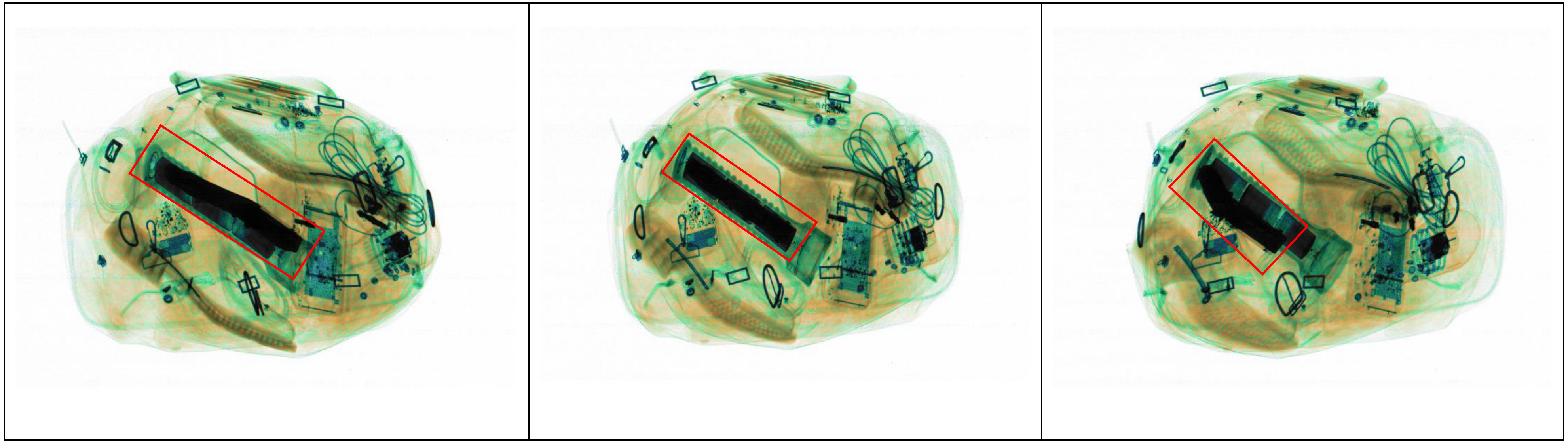
Bag 15: Small size kitchen knife partially masked by a laptop, in the periphery of a canvas bag.



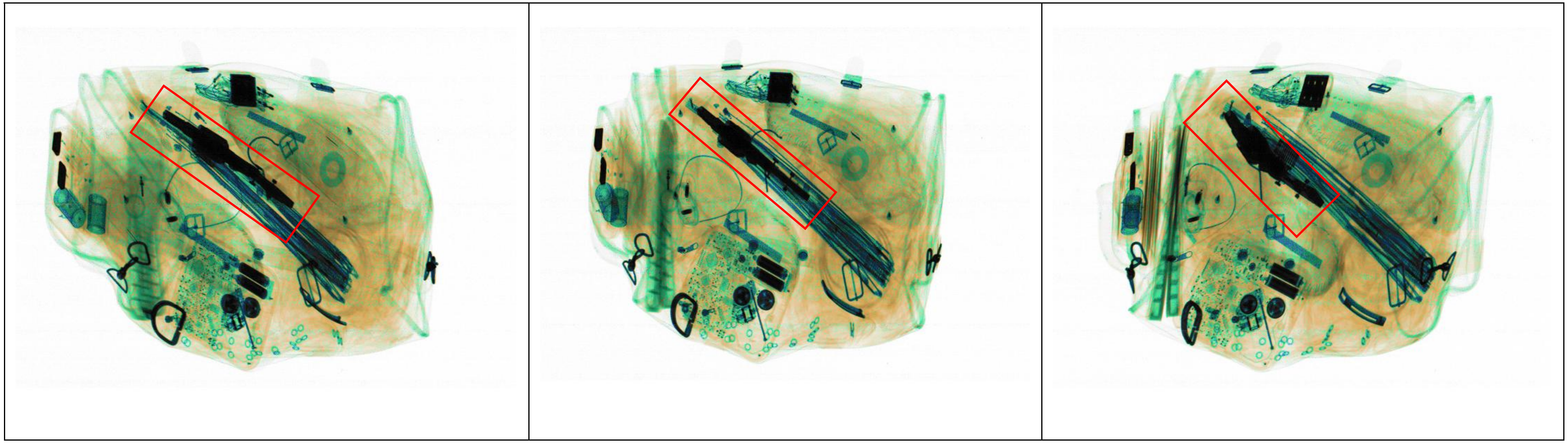
Bag 16: Small size kitchen knife (ceramic) at a minimum footprint with no mask, in the periphery of a small soft case bag.



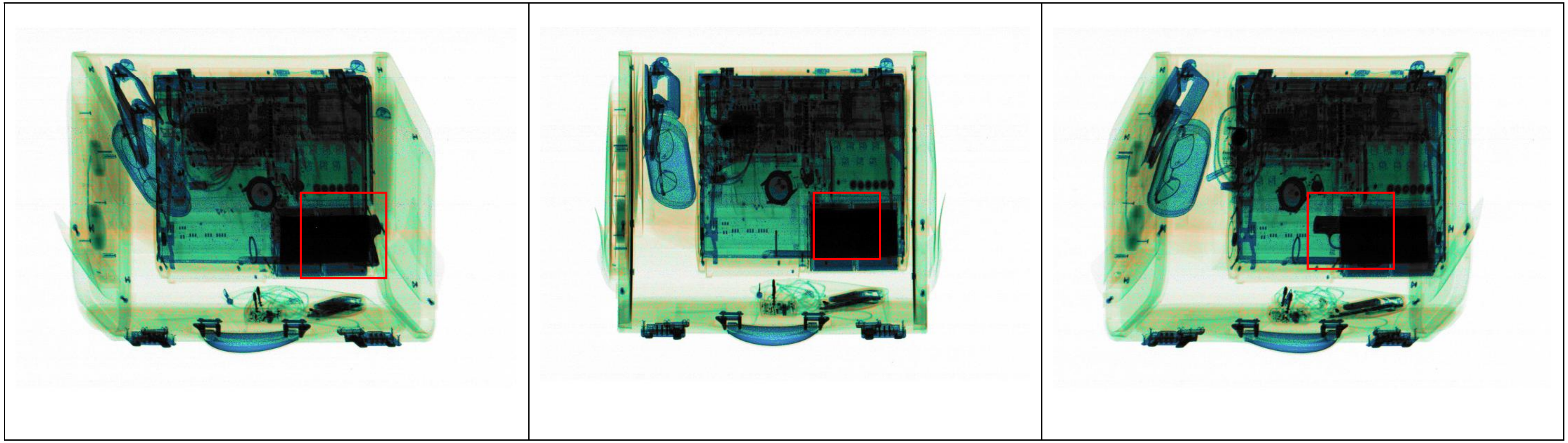
Bag 17: Large bread knife masked by a curling iron and a flashlight, in the centre of a soft case backpack.



Bag 18: Medium size pistol (Beretta) completely masked by a flashlight, in the centre of a vinyl bag.

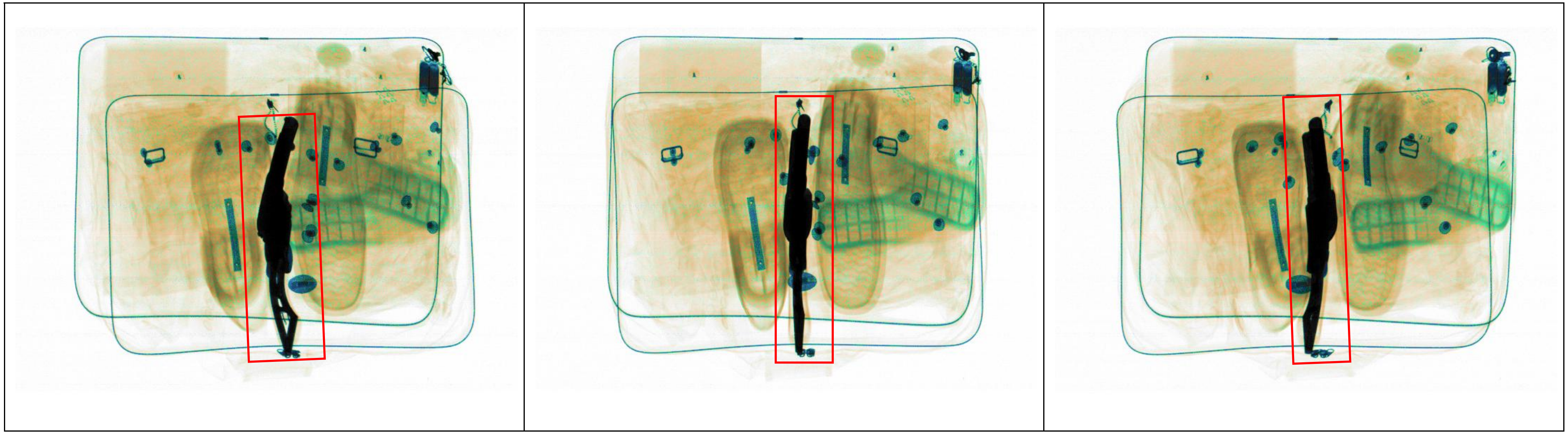


Bag 19: Medium size revolver (S&W .38 Special) masked by an umbrella, in the centre of a canvas bag.

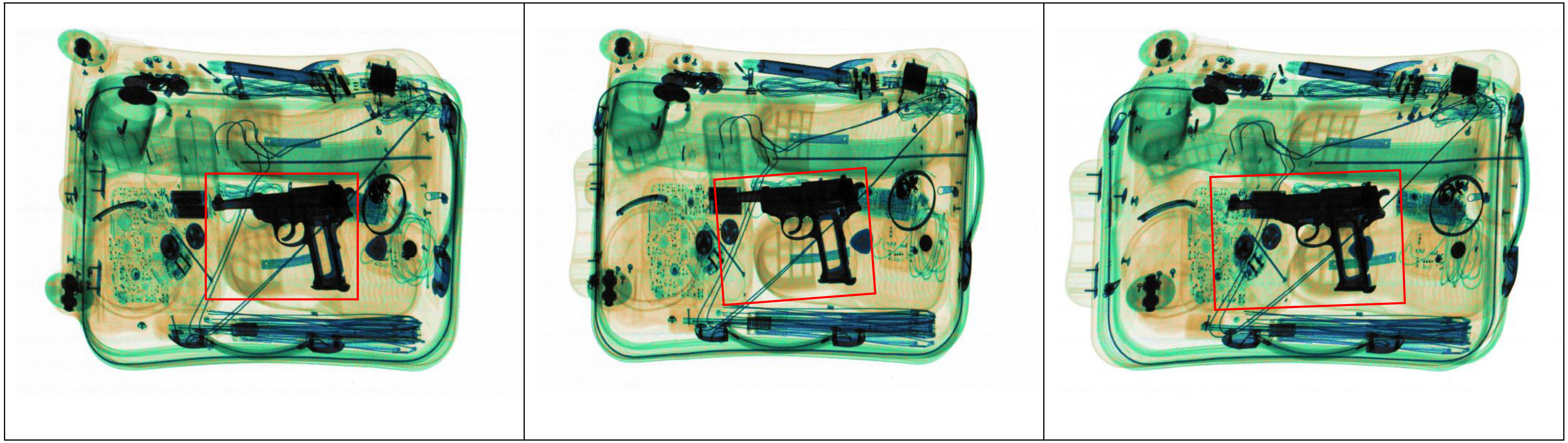


Bag 20: Small pistol (Baby FN) masked by a laptop, at the periphery of a leather briefcase.

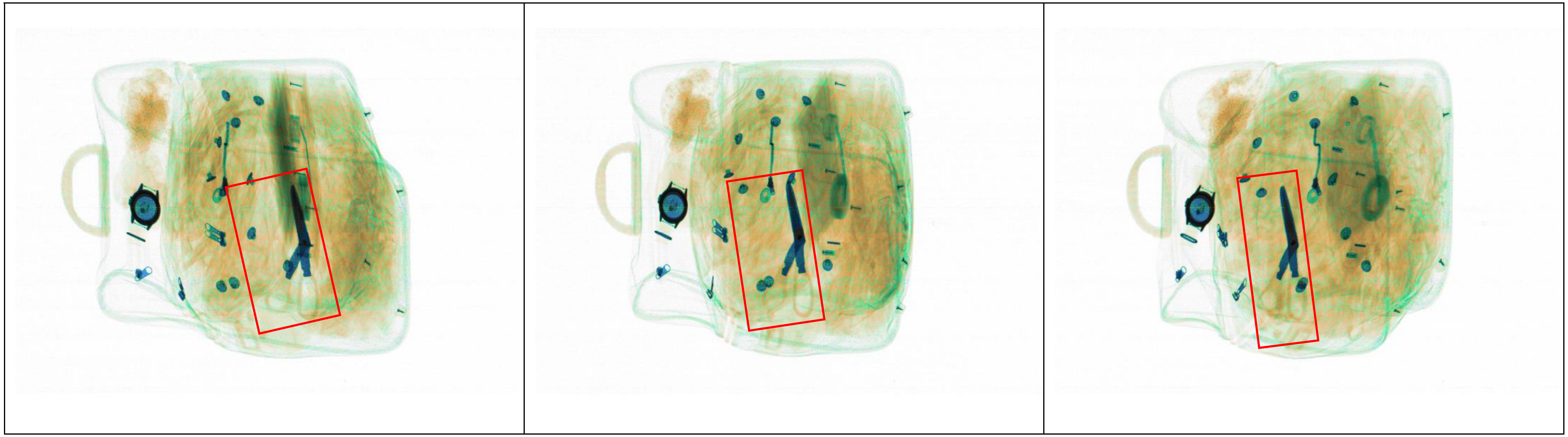
Figure C-2. Training Images -20° (Frame 1), 0° (Frame 21), and +20° (Frame 41) views. The target is marked by the red rectangle.



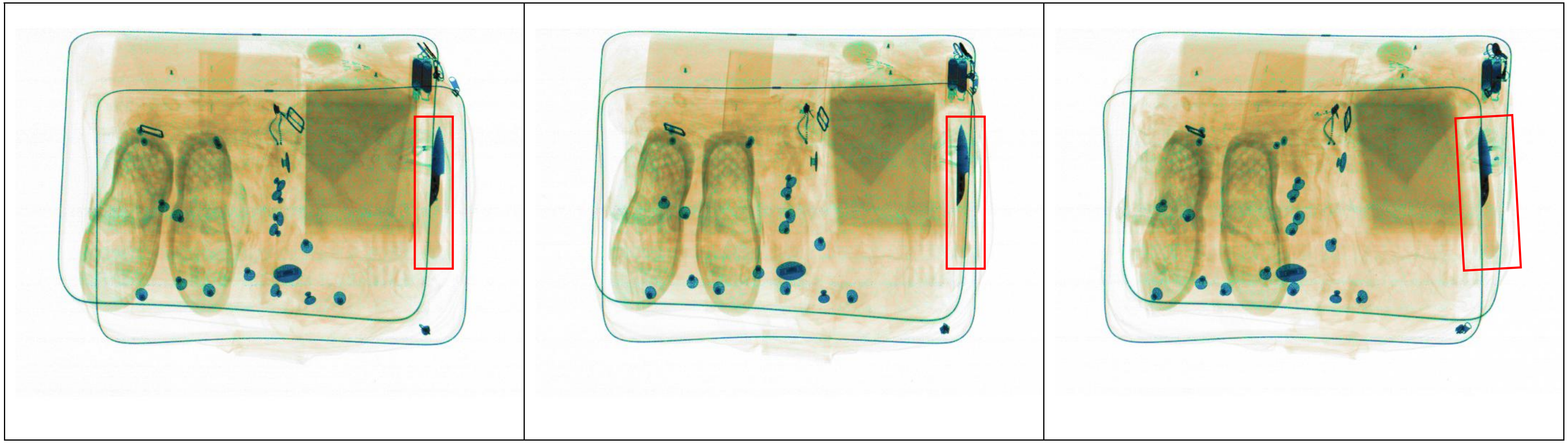
Bag 1: Large revolver (Colt SSA .45 Cal) with no mask but at a minimum footprint, in the centre of a canvas bag.



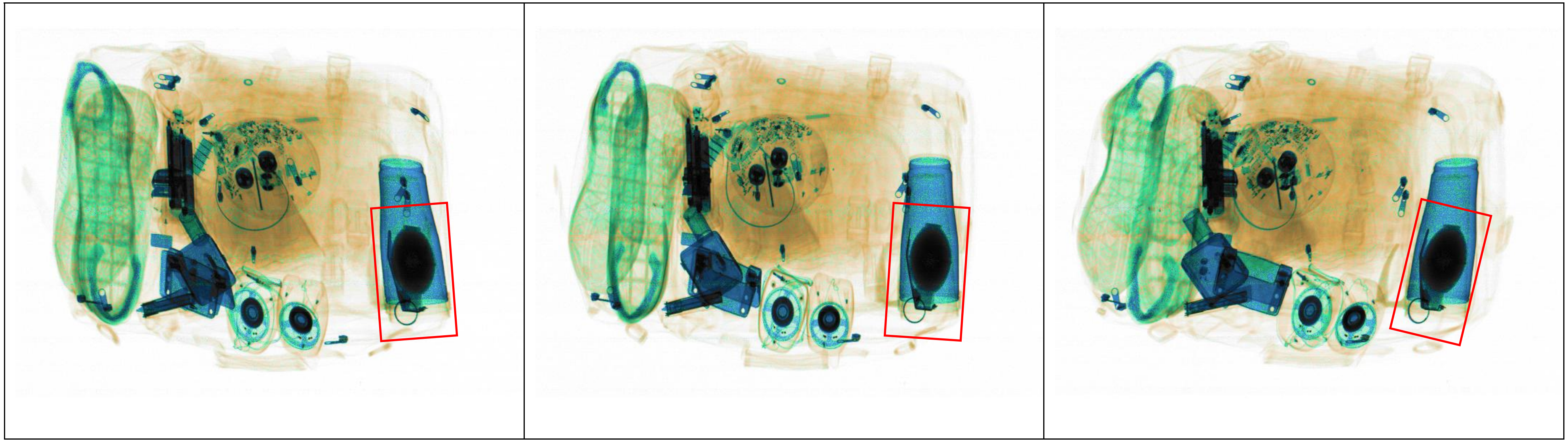
Bag 2: Large pistol (Walther P38) unmasked and at maximum footprint, in the centre of a wheeled bag.



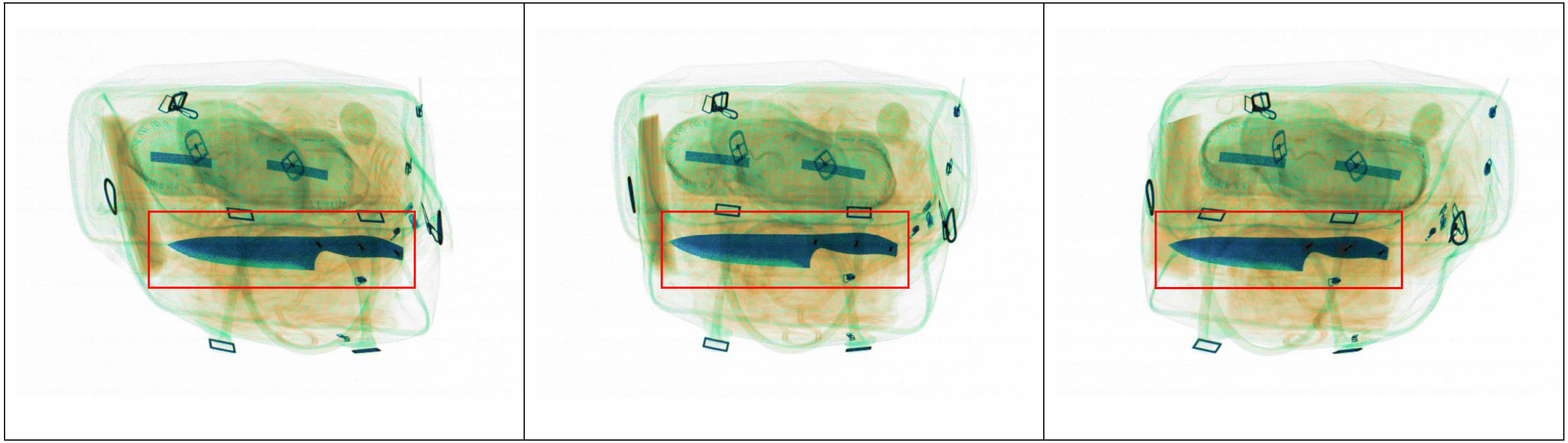
Bag 3: Large scissors with plastic handles no mask and maximum footprint, in the centre of a small soft case bag.



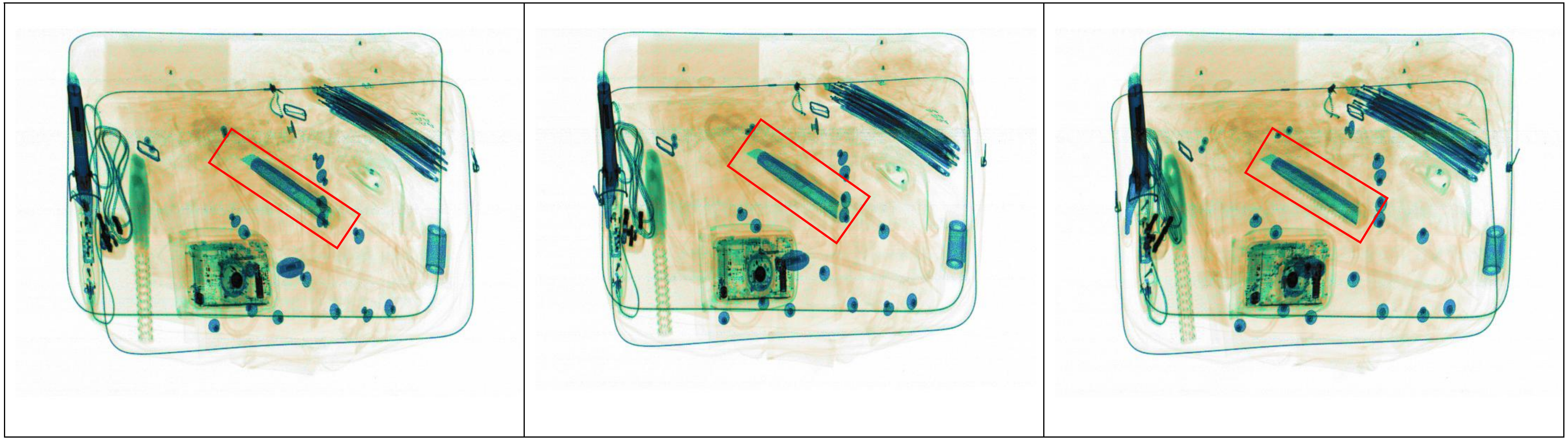
Bag 4: Small kitchen knife (ceramic) with no mask and at a maximum footprint, at the periphery of a canvas bag.



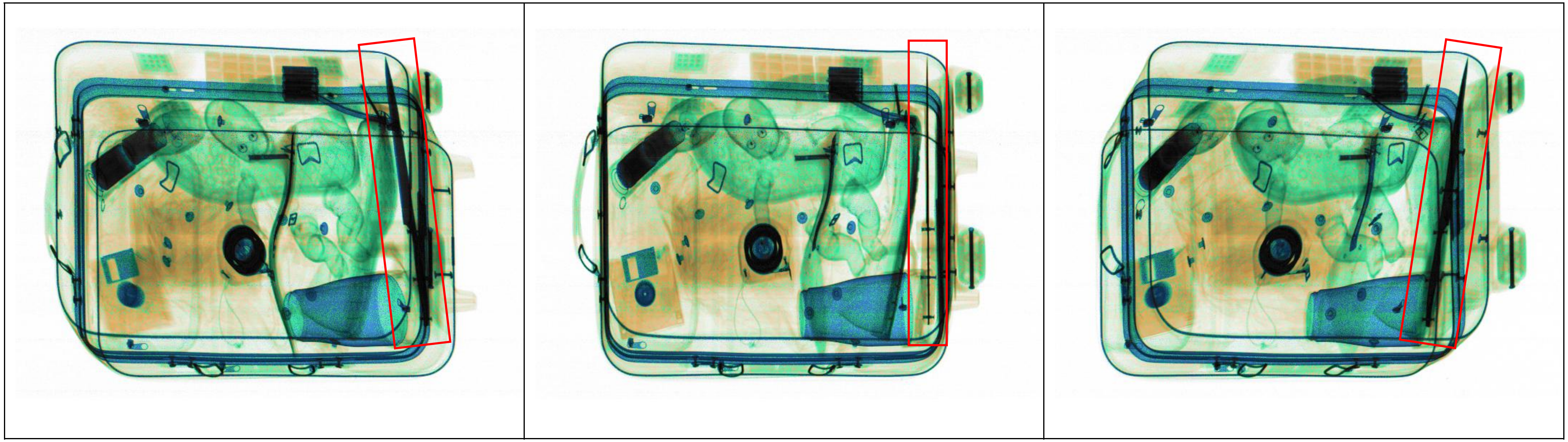
Bag 5: Lemon grenade (solid metal used for training) at a maximum footprint with no mask, at the periphery of a soft case backpack.



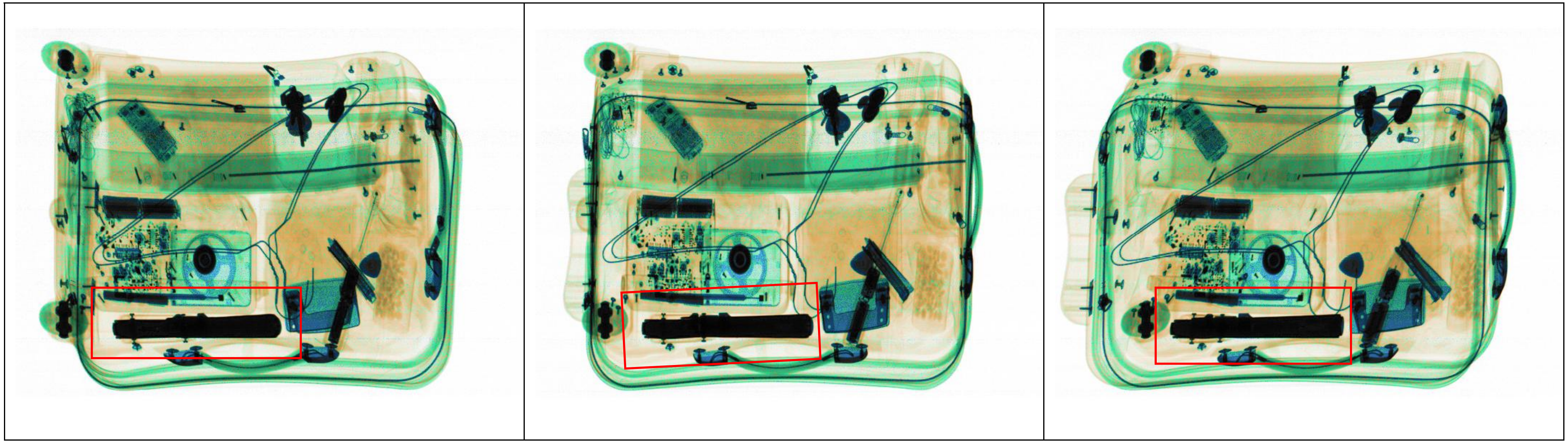
Bag 6: Large chef's knife at a maximum footprint and with no mask, in the centre of a vinyl bag.



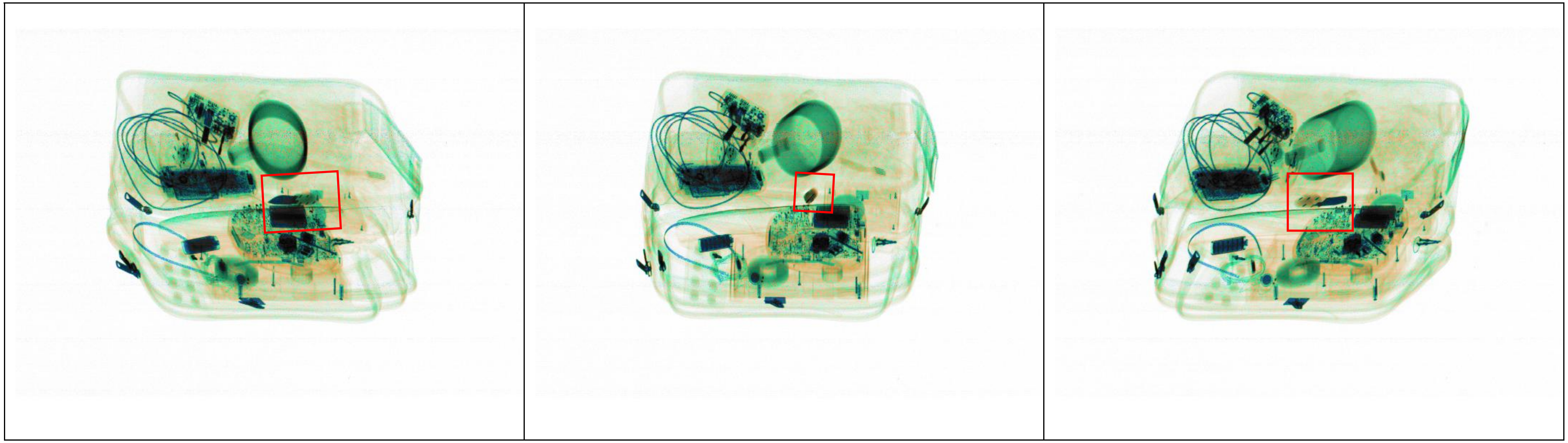
Bag 7: Box cutter (utility knife) at a maximum footprint (blade extended) and with no mask, in the centre of a canvas bag.



Bag 8: Large chef's knife at a minimum footprint but with no mask, in the periphery of a wheeled bag.



Bag 9: Large pistol (Colt 1911 .45 Cal) at a minimum footprint but with no mask, at the periphery of a wheeled bag.



Bag 10: Medium kitchen knife at a minimum footprint but with no mask, in the centre of a small soft case bag.

List of Publications

Evans, J. P. O., Chan, J. W., Vassiliades, V, Downes, D., & Kaufman, L. (2006). Kinetic depth x-ray imaging (KDEX) for security screening. *Proceedings of the 4th International Aviation Security Symposium*, Washington DC, November/December 2006, 313-319.

Evans, P., Vassiliades, V., Chan, J. W., Kaufman, L., & Downes, D. (2007). Human factors evaluation of kinetic depth x-ray imagery for aviation security screening, *Contemporary Ergonomics* (Taylor and Francis, London), 190-196.

Vassiliades, V., Evans, P., Kaufman, L., Chan, J. W., & Downes, D. (2008). A human-centered approach for improving airport security. *Contemporary Ergonomics* (Taylor and Francis, London), 127-132.

Vassiliades, V., Evans, P., Kaufman, L., Chan, J. W., & Downes, D. (2008). Preliminary evaluation of kinetic depth x-ray imagery for aviation security screening. *Cognitive Technology*, 13 (1), 18-25.