Automatic Surveillance and CCTV Operator Workload

Submitted September 2008, in partial fulfilment of the conditions of the award of the degree M.Sc. in Interactive Systems Design

Nastaran Dadashi

School of Computer Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature _______________________

Date _____/_____/_____

[I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via the University of Nottingham's e-dissertation archive.]*

Abstract

Despite the recent technological advancements in developing automatic surveillance systems, their effectiveness is usually affected by different factors and they are not completely reliable. These "imperfect" systems might still be useful if they inform users of their deficiency. In other words, if the automatic system could inform the user about its level of confidence in the generated alarms, then the users' reliance on the system would be increased. A series of experiments were designed to investigate various design features such as functionality of automatic system (Spotting and Tracking) and different levels of accuracy (66% and 75%). Also, three levels of "confidence information" (no confidence information, unreliable confidence information and reliable confidence information) were introduced which demonstrated the likelihood of the alarm to be true positive. The workload of operators was measured to investigate the efficiency of each of the modes. A total of 24 students took part in these experiments. The results showed significant reduction in workload when there is "reliable confidence information". Slight increase in the accuracy and variation in functionality failed to produce evidence of change in workload. This project confirms that effectiveness of imperfect automatic surveillance system currently available can be improved with simple non technical design modifications. This project also emphasises the need for further research in the area.

Keywords: Automatic surveillance systems, CCTV, workload assessment, reliance

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgement

First of all, I would like to express my gratitude to my supervisors Dr. Tony Pridmore and Dr Alex Stedmon for their help, fruitful discussions, insightful advice, enthusiasm and encouragement throughout my research. Their efforts in helping me in the development of the project, through technical difficulties and in search for relevant literature are much appreciated.

Finally, I am always indebted to my family, especially my parents. I would like to thank them for their love, support and encouragement.

1 Introduction

Closed Circuit Televisions (CCTV) have long been used as a powerful surveillance tool for fighting crimes. As Ainsworth (2002) pointed out, CCTV is the finest weapon in security armoury. Video security and surveillance technology has advanced further and faster in recent years than any other time period. However the effectiveness of these systems is still mostly limited to post-incident investigation. Automatic surveillance systems are developed to expand the effectiveness of CCTV cameras, but the extent of their effectiveness depends on various factors in addition to the technological advancements.

Automation is used in complex situations to reduce the workload put on the human operator and consequently to increase the productivity and enhance the performance. Workload assessment of operators while using an automatic surveillance system can provide system developers with information about the systems' effectiveness. Furthermore, depending on automation is not only related to the accuracy of the system but also to the human operators' reliance on automatic system

The design of automatic surveillance systems can determine the level of reliability and increase the effectiveness of the surveillance. Automatic surveillance systems are interactive systems and human operators play a central part. Therefore, applying principles of Interaction design should theoretically enhance the performance.

"Offer informative feedback" is one of the eight golden rules of interface and interactive system design proposed by Shneiderman (Shneiderman 1997). The feedback given to participants in this study was the level of systems' confidence in the generated alarms to be true positive. The idea was similar to a study conducted by Dzindolet et al., (2003), where users of an automatic aid could see a bar graph of their own performance and the aid's accuracy. It was assumed that giving users some feedback regarding the performance of automatic aid can increase the trust and reliance on the automatic system. Similarly if CCTV operators were informed about the systems' confidence in detecting a target, their performance should presumably increase.

Riley (1996), proposed a theory for automation use. In this theory a set of factors affecting the reliance on automation was identified. These factors were: risk, fatigue, state learning, confidence and trust. In a similar study by Jamieson (2007), usability of the automatic aid was considered to influence the users' reliance on the system.

Automatic surveillance systems are used for detecting objects and tracking them and in higher levels to analyse the behaviour of targets (Valera and Velastin, 2005). In operator's surveillance both detecting and tracking are bound together (i.e. one cannot track a target without detecting it first). However in automatic surveillance system there might be a trade-off between the two. Cases that are easy to detect are not necessarily easy to track and vice versa (Dick and Brooks, 2004). It is useful to know which one is easier for human operator to enable automatic system developer to choose when compromising is necessary.

Although computer vision algorithms responsible for detection and tracking have reasonably acceptable performance in laboratory, in real life situations their effectiveness is no where near acceptable (Dee and Velastin, 2007). In this project an automatic surveillance system was prototyped using the "Wizard of Oz" technique. Two levels of accuracy have been studied, the first set (66%) was obtained from the similar systems currently available in the literature and the second one (75%) tested the effects of a small increase to the accuracy.

The automated system simulated in this project provides users with information termed in this report as "confidence information". This information is the extent to which the system is certain about its choice. Furthermore two system functionalities (detection or tracking) and difference accuracy levels (66% and 75%) were investigated.

This report consists of 6 chapters. Chapter two reviews the published work and related research. This information has been used as the base for the experiment designed in this project. Chapter three describes the methodological strategy applied for the implementation of the experiment and the apparatus required for setting up the experiment. Chapter four reports the results of the experiment and chapter five discusses the results and justifies them as well as comparing them with similar studies. Finally, chapter six attempts to derive a set of recommendations for future studies.

2 Literature Review

In this section, the published work on monitoring Closed Circuit Television (CCTV) and automatic surveillance systems has been reviewed. In section 2.1, the general literature on Closed Circuit Television has been reviewed. Section 2.2 covered CCTV operators' responsibilities and the challenges which imply the need for automatic systems which are reviewed in Section 2.3. Section 2.4.discussed the features of current automatic surveillance systems and their effectiveness. Section 2.5 studied the human automation interaction and the factors affecting users' reliance on automation .In order to assess the effect of difference design features of automatic surveillance system, workload assessment techniques was reviewed in Section 2.6.

2.1 Closed Circuit Television

2.1.1 Overview

We live in a "Big Brothers'" world. According to "Nacro", Crime and Social Policy Section briefing (Armitage, 2002) the extent of Closed Circuit Televisions (CCTV) coverage in the UK is estimated by 5,238 cameras in 1997. This number has increased dramatically to 4,000,000 cameras in 2007 (Armitage, 2002). The aim of applying CCTV is to reduce social offences and crimes by deterrence, monitoring the scenes (reactive surveillance), alteration of a potential victim, presence of a capable guardian and detection (proactive surveillance) (Armitage, 2002).

Unlike robbery and assault which can be reduced through the deterrence effect of CCTV and might be solved through forensic analysis of video footages, public transport systems such as the rail way stations are more vulnerable to severe events such as terrorist attacks, suicidal attempts and trespassing which require continuous monitoring of video footages. Automatic surveillance systems are to assist operators in this matter.

The present project aims to study the effectiveness of automatic surveillance systems and the extent in which it can reduce CCTV operators' workload. Another issue affecting the performance of any automatic system is operators' reliance on the system. This issue has also been addressed in the present project.

2.2 Monitoring CCTV cameras

CCTV has been widely used in recent years. In time they have changed from a merely flow control system responsible for general monitoring to individual surveillance system assigned with suspect detection (Cameron, 2004). A CCTV system consists of several cameras, monitoring and recording systems and control room operations. A CCTV operator can monitor CCTV displays either reactively or proactively. This section focuses on the surveillance task and the responsibilities of a CCTV operator.

2.2.1 Proactive monitoring

Proactive surveillance refers to online monitoring of the scenes to detect and predict suspicious events and prevent incidents from happening. Despite the importance of proactive surveillance, the scope of proactive surveillance carried out in reality is very little. In a field study performed as part of the URBANEYE project by (Norris and McCahill, 2006), CCTV operators were observed while performing their tasks. The results revealed that only 35% of the total detection was due to proactive monitoring. Furthermore, according to Velastin et al., (2006), after 20-40 minutes of active monitoring, CCTV operators will suffer from "video-blindness", and cannot recognise objects on the video anymore.

Proactive monitoring is a demanding task and imposes a high level of workload upon the controller which makes the whole surveillance task even more difficult. Based on a study published in Security Oz magazine:

"After 12 minutes of continuous video monitoring an operator will often miss up to 45% of screen activity, after 22 minutes of viewing, up to 95% is overlooked. " (Ainsworth, 2002, pp.20)

Even if the proactive monitoring is as part of the CCTV operators' responsibilities, it is very demanding and impossible for a long time. Thus the dramatic increase in the number of CCTV cameras can not guarantee increase in security.

2.2.2 Reactive Monitoring

Reactive surveillance is passive monitoring of CCTV cameras after being alerted by audio feedback. Keval (2006) performed a series of ethnographic observations and semi-structured interviews in six CCTV public surveillance rooms; five of them in London and one outside London. It is found that reactive surveillance is the most heavily used surveillance task which is usually triggered by police and business radios.

Another interesting issue from Keval's study (2006) is that the majority of security surveillance task is audio driven rather than video driven. Luff and Heath (2001) reported a typical day of controllers and their routine responsibilities. Based on this report operators don't monitor every single screen at all times. When there is a report about some suspicious behaviour on the station, the station supervisor finds the suspect on the screens and tracks the target. After locating the target and studying the situation the operator informs an on-site security staff to investigate the situation in person.

2.2.3 Responsibilities

CCTV operators can control up to 16 cameras simultaneously. This number varies based on the complexity of the scene. Interview with operators conducted by Neil et al. (2007) revealed that operators are confident when they simultaneously monitor four screens. Moreover, to keep up with the health and safety guidelines an operator should take a break every 20 to30 minutes (Wallace and Diffley, 1998).

Note must be taken that monitoring CCTV cameras are not the only task assigned to an operator. The observational study by Luff and Heath (2001) in London Underground CCTV control rooms showed that a CCTV operator is responsible for a number of other tasks. These tasks range from responding to passengers' complaints to updating the staff about any problem which has occurred on the network or any other critical incidents. Hence, the operator is an integral part of the controlling system, it is neither merely the technology nor a set of regulations that runs the control room; it is the operator (Donald, 1999).

2.3 Challenges in monitoring CCTV cameras

2.3.1 Situation Awareness

Situation awareness is a phenomenon associated with tasks in dynamic environments. It is one of the important factors in effectiveness of human-machine interactions, originally used in aircraft pilot literature. Situation awareness is a cognitive definition of operators understanding and reactions towards their environment:

"Situation awareness is the perception of the elements in the environment within a volume of time and space (Level 1), the comprehension of their meaning (Level 2), and the projection of their status in the near future (Level 3)." Endsley (1995, pp.36)

In the context of CCTV monitoring, an operator should perceive a number of factors such as environmental conditions, location and characteristics of the area in different times. Operators have to interpret the information obtained in the first level of situation awareness theory. For instance, they need to decide whether the pattern of crowd behaviour is suspicious. Finally after interpreting the perceived information (level 2), operators need to make decisions. Whether they will send someone on site to check the abnormality or leave the event because it does not seem to be critical (level 3).

Due to the complex nature of human behaviour and the large number of cameras which should be monitored, it seems that maintaining operators' situation awareness is very challenging.

2.3.2 Vigilance

Vigilance for CCTV operators refers to the capacity of sustaining effective attention for monitoring CCTV screens and remaining alert for long periods. Vigilance is one of the reasons for the limited capacity of the proactive analysis of CCTV cameras, especially due to its pervasive nature.

Since operators need to deal with several tasks simultaneously, the performance will be inevitably affected. Donald (2001) describes the vigilance in monitoring tasks and its affect on the overall performance of the operators:

"Vigilance intensive operations have as their main challenge a requirements of consistently accurate attention and detection, and the maintenance of performance levels under these conditions….a CCTV operator may never be aware that something has been missed(Donald, 2001,pp.37)" .

In other words, the quality of sustained attention deteriorates in time. As mentioned earlier remaining vigilant can cause video-blindness. Furthermore, there is a risk of missing events while the operator is attending another suspicious situation.

2.3.3 Mental Model

Images obtained from CCTV cameras represent spatial information. Cameras are shown by numbers, and each is associated with a location in the whole area under CCTV coverage. CCTV controllers have to know the area under control. However, due to the high number of cameras installed, it is hard to remember the exact location and this will reduce the high response rate required in critical situations. Keval (2006) noted that:

"Operators regularly shouted to colleagues across the room if they were stuck and could not recall the camera number or its location. Operators shouted louder and in a panic-like tone particularly when communicating with police operators via telephone or radio to follow targets of interest onscreen."(Keval, 2006, pp3).

Operators' mental model of the location is built in time. An expert operator have better mental model of the area under coverage than a novice worker. However, in highly hazardous situations where decisions should be made very quickly, temporal demand and stress might decrease the effect of expertise. As the case described in (Keval 2006), relying on expertise is not a guaranteed solution.

General aspects of CCTV monitoring have been covered in this section. Challenges mentioned emphasise the necessity of implementing automatic surveillance systems. The next section will briefly explain the current automatic surveillance systems in the literature.

2.4 Automatic surveillance systems

This section discusses various aspects of automatic surveillance systems, their main features, and the core challenges towards their effectiveness. Automatic surveillance systems monitor the displays and alert the security controller when there is an extraordinary situation. In other words they provide automatic real time interpretation of scenes to the operators. "*The main goal of automatic surveillance systems is to obtain a description of what is happening in a monitored area and then to take appropriate action based on that interpretation."(Dick and Brooks, 2004, page 160)*.

Automatic systems are inevitable when:

1- The task is impossible to be done manually, Proactive surveillance seem impossible since "*there are too many cameras and too few pairs of eyes to keep track of them". (Hogan, New Scientist, 2003, page 4)*.

2- The task is difficult or unpleasant to be done by human operators, for example detecting unattended bag in a university campus (Harris et al., 2008).

3- Automation can extend human capability, for example response time would increase with the aid of automatic surveillance systems.

Despite the recent improvements in developing intelligent surveillance systems, their effectiveness is still not sufficient to meet real-time surveillance requirements. This will be discussed in more detail in the following section.

Moreover, apart from effectiveness of automatic surveillance system, there are factors which are responsible for acceptance of such automated systems by operators which are classified in (Mosier and Skitka, 1996):

1- The extent to which organisations encourage the use of automated systems (acceptability)

2- The extent to which reliance on automated decision aids is acceptable

3- The extent to which automation is seen as an expert and one that is perceived to yield more accurate and reliance outcome than people's own judgment.

2.4.1 Automatic surveillance systems' key functions

Automatic surveillance system is mainly designed to help operators in continuous monitoring and its main function is to support proactive monitoring. However, these functions can also be applied in reactive monitoring if the operator could be able to switch to the automated mode after being alerted.

As noted in a review of intelligent surveillance systems by Valera and Velastin, (2005), the main processing stages in automatic systems are object detections and recognition, tracking and behavioural analysis (Sage and Young, 1999; Davies et al., 1995). Valera and Velastin, (2005) suggest figure 2.1 which indicates the main processing modules in an intelligent surveillance system:

Figure 2-1: Main processing modules in automatic surveillance system (pp. 194)

The first three module, object detection and object recognition (Sanderson et al., 2007; McKenna and Gong, 1997) and tracking (Hayman, 2003; Moeslund and Granum, 2001) have received a lot of attention in recent years in the computer vision research community.

Many automatic surveillance systems have been designed for public transport systems. In Valera and Velastin (2005), there are 12 state of the art third generation surveillance systems mentioned. However as Dee and Velastin (2007) pointed out:

"*with the significant challenge that these places(i.e. public transport facilities) are subject to high crowding levels, conditions in which people tracking is well beyond the current state of the art."(Dee and Velastin, 2007, pp. 8)*

Third generation surveillance systems is the term used in Valera and Velastine (2005) to refer to the large distributed networked cameras with various monitoring points to address complex situations of human monitoring processes. This group of automatic surveillance systems tend to consider the requirements of end-users who in this case are security operators.

Some of the most promising commercial automatic surveillance projects are DETEC (www.detec.no) DETER, CROMATICA, and PRASMATICA. PRASMATICA is a project funded by EU (Lo et al., 2003) which was an extension to another EU funded project called CROMATICA. The latter considered various inputs such as data from the CCTV cameras, local wireless cameras, smart cards and audio sensors. An intelligent system defined in PRASMATICA networked this information to the central control unit. Moreover, PRASMATICA took into account the legal, ethical and sociological effects of introducing such distributed systems. It seems that this final project should solve the automation problem of surveillance systems; however the centralized approach towards various sources of information caused deficiencies and decreased the performance of the system.

2.4.2 Challenges

Challenges for developing an effective automatic surveillance system can be categorised in two groups: 1- general vision problems and 2- the features of the surveillance environment. The core application of every automatic surveillance system is to detect objects and track them, classify objects and analyse behaviour (Dick and Brooks, 2004). Furthermore, Dee and Velastin (2007) mentioned that robustness is also an important issue in developing automatic surveillance systems. A robust system is one which can function reasonably in stressful environments. On the other hand, most of the automatic surveillance systems are domain specific and can hardly be generalised to other environments (Dick and Brooks, 2004). The obstacles for having robust automatic surveillance systems as listed in Boghossian and Black (2005) are:

- Scene illumination

  Illumination of the environment varies with time, and different weather conditions. These variations change the observational characteristics for the automatic surveillance system and decrease the performance of the system.

- Movement of background objects

Fixed items in background such as trees might also move which confuses the automatic system and increase the false alarms.

- Long term variations

Background also changes due to change in season, this increase the false alarms unless the automatic system is designed to gradually adapt with the background alterations.

- High density scenes and clutter

This is the main issue in public transport surveillance systems, and it is related to highly crowded places. High density will lead to occlusion which as mentioned by (Dee and Velastin, 2007) is the most important problem facing tracking.

To deal with these challenges automatic surveillance systems tend to set a threshold, in other words this threshold is the level of sensitivity of the system, if the system is highly sensitive which have a low threshold it will generate too many false alarms, and if the threshold is set too high which makes the system's sensitivity low it will miss suspicious events.

2.5 Automation and quality of surveillance

Although automatic surveillance systems are inevitable especially with the increasing necessity for security systems in the recent years, they are not as effective as they should be. Riley (1996) proposed an initial theory of operator reliance on automation which was investigated through a set of experiments on a simple automated system (a computer game). The results led to a revised theory of automation use, which is the basis of the current project, (Figure 2.2). Dashed lines represent influences which have not proved to be significant and straight lines illustrates significantly influential issues which have an effect on the use of automation by human operators.

In this project, workload, task complexity, accuracy and reliance from Riley's theory of automation use have been investigated. Moreover, a set of experiments conducted by Jamieson (2007) found another aspect which has an impact on automation which is effectiveness.

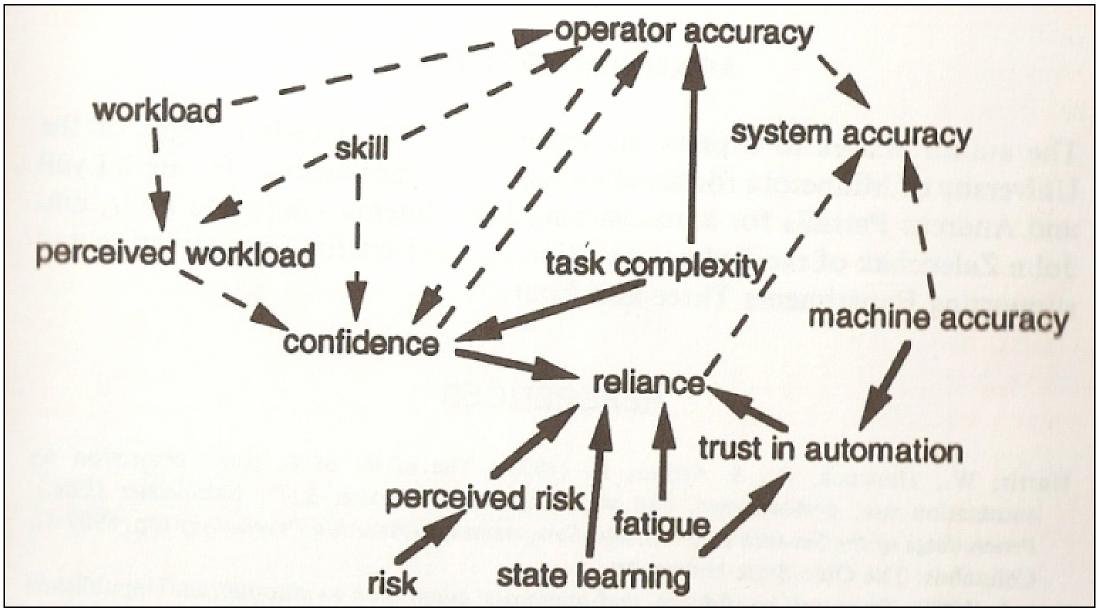


Figure 2-2:Revised theory of automation use (Riley, 1996, pp.33)

2.5.1 Workload

Mental workload (MWL) is defined as (Sanders and McCormick, 1987, pp.69):

"*The difference between the amount of resources available within a person and the amount of resources demanded by the task situations*."

In other words MWL is the cost put upon the operators' information processing capacity while monitoring CCTV cameras. Excessive demand on their available resources will lead to overload and performance degradation. Stanton et al., (2005), considered MWL as a combination of interacting stressors on an individual, the framework shown in Figure 2.3.is taken from (Megaw, 2005)

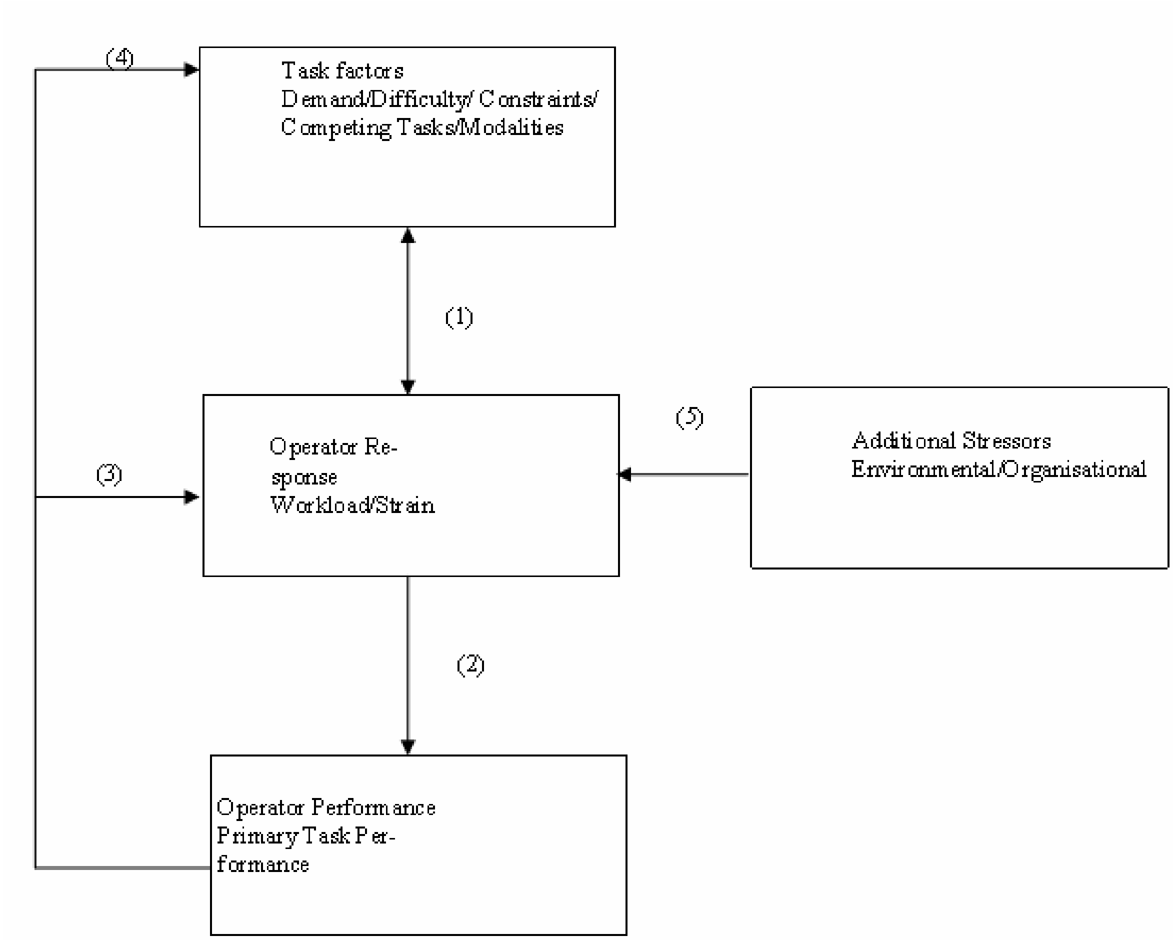


Figure 2-3: Factors of operators' workload (Megaw, 2005, pp. 526)

This diagram suggests that workload depends on the different characteristics of the task where quantity of work is just one of these factors. External stressors such as social pressure on CCTV operators to detect crimes or the time and location the surveillance is taking place can also affect the workload. These factors define different organizational dimensions to the surveillance task (Megaw, 2005, pp. 526):

"Operator workload is not simply a function of task demand but is influenced by how the task is perceived by the operator, sometimes referred to as cognitive appraisal"

Here is an example to elaborate this diagram for surveillance task. Operators' workload depends on a number of factors including the surveillance task and its constraints. This explains the two way relation in (1). External stressors such as the time and location of surveillance as well as operators self confidence also affect the workload (5 and 3).For instance, if an operator expects himself to proactively monitor all 16 screens in a control room without making any error, potential stress will increase the workload. In case of a large number of false alarms, the operator has to spend more time and energy to handle it and make sure that while he is busy dealing with the false detection, he has not missed any real suspicious event on the screens(4).

Assessing mental workload to identify the level of available resources is a very important step towards sustaining a reasonable performance.

2.5.2 Task Complexity

One of the problems with automation is that it automates easy tasks and leaves the hard ones to the human operator and makes the automatic system "clumsy" (Kantowitz and Campbell, 1996). It is important to understand the complexity of task to know which one should be automated to avoid clumsiness.

Detecting suspicious behaviour and tracking are the two key functions of any automated surveillance systems (Ferryman et al. 2008). The main purpose of detection is to find something suspicious, moreover the target has to be tracked and the operator should be able to predict the next location of the target.

There is a trade off between object detection and tracking. There are cases which are easy to detect but hard to track, such as features around the corners. On the other hand there are features which are hard to detect but easy to track, such as very complex shapes (Dick and Brooks, 2004). Thus it is important to understand which of the automated functions has higher performance and lower workload.

2.5.3 Accuracy

The reliability and robustness of automatic surveillance systems is one of the factors that have to be considered for development of successful automatic surveillance systems

(Muller-Schneider et al., 2005). Although the projects reviewed in this section are promising, they are still far from a real time perfect intelligent monitoring system due to the challenges mentioned earlier. The purpose of this section is to find the accuracy of current automatic surveillance systems and apply the mean percentage of that accuracy to the experimental system mocked up for the purpose of this project.

Tracking is a major part of current automatic surveillance systems; object identification and motion tracking have been investigated for more than two decades. McKenna and Gong (1997) developed a system for detecting and tracking several people (multi-track) as they move through complex scenes. In their system, a motion based tracker which stored a history of each pixel and clustered them into various areas of motion to differentiate one person from another, as well as a face tracking system from each frame to another has been used to provide information for the motion clustering process.

A systematic method to evaluate the performance of tracking algorithms suggested by PETS (Performance Evaluation of Tracking and Surveillance) which was developed through a series of workshops started from 2000. PETS has been a used as milestone for evaluating tracking algorithms. (www.petsmetrics.net).

Muller-Schneiders et al., (2005) evaluated their proposed algorithm against PETS dataset and the quantitative results they got for tracking true positive was 85%. True positive is the case when the system correctly detects and tracks an object in relation to the ground truth data. False positive is when the target is being detected but there is no reference to it in ground truth data and false negative is missing an event from a ground truth data. In another automatic surveillance system proposed specifically for railway stations by Black, et al., (2005), the percentage of true positive rate for tracking an object within the defined range of interest was 76% (Table 2.1). The green column in Table 2-1 represents the accuracy of true positive detections (TP) for various cameras tested in Black (Velastin and Boghossian, 2005).

Camera	TP	FP	FN
A	75%(9)	25%(4)	25%(4)
C-IR	77%(10)	0%(0)	23%(3)
C	77%(10)	0%(0)	23%(3)

Table 2-1: Accuracy of tracking system in Black, Velastin and Boghossian(2005), pp. 194

Target detection in automatic surveillance systems is a very useful and yet challenging task. It is useful because it can be more accurate than a human operator and it leads to faster reaction to suspicious situations. On the other hand, various illumination conditions, facial expressions and pose angles can make target detection a very challenging task.

There are many algorithms implemented for face recognition systems. Moreover to get an idea of effectiveness of the algorithms, FERET test protocol was developed (http://www.frvt.org/) FERET evaluate algorithms and asses their quality. Sanderson et al., (2007) employed an intelligent CCTV for "face in the crowd" recognition through an approach called Bag-of-Feature. In this approach the face is considered as a set of featured vectors, these vectors were obtained from clustering the face into blocks. A bias was added to each vector to consider the illumination variations and finally the probability of vector 'x' being a part of block 'i' is taken into account.

Ideal face recognition is the one which accurately detects the object with least false detection. Hampapur et al. (2007) explored various video parsing techniques to improve searching suspicious events within the surveillance system. The accuracy of the performance for tracking objects in this approach was measured 73.6 % and 78.8 % for object detection task. The green lines in Table 2.2 is the rate of misses in false negative.

Object Detection or Background	Subtraction Results
False Positive	0.03 objects per frame
False Negative	628 of 2964=21.2%
Avg size of missed object	226 sq-pixels

Object Tracking Results	
False Positive(spurious tracks)	25 with average length of 77 frames
False Negative(missed tracks)	24/90=26.6%
Avg size of missed object	169 sq-pixels

Table 2-2: Object detection performance and object tracking summary from Hampapur et al., (2007) ,pp. 80

Another system proposed by Shan et al. (2007) aimed to deal with various illuminations and pose angles through a pose variability compensation technique. The idea was to identify face through its pose and the correlation of that pose against the database which was available to the system. The result they got based on the FERET test protocol had the accuracy of 57%. None of the systems studied in this review are 100% accurate, the question is how accuracy can affect performance and workload of CCTV operators and how increase in accuracy from one imperfect automated system to another can dramatically change the quality of surveillance.

2.5.4 Trust in automation

Automation reliability is defined as "*the extent that automation can be said to be reliable and does what the human operator expect it to do*" (Wickens et al., 2004, pp.423). This is not only dependent on the accuracy of the automatic system. If the automatic system provides users with some information about the extent of its accuracy, users' trust will also increase.

In a study conducted by Dzindolet et al., (2003), it is mentioned that after observing automated aids making errors, operators lose trust in the system. On the other hand, Dzindolet et al. (2003) noted that when the user knows why and when the automated system might make mistakes, then the user reliance on the automated system will increase even though the automated system is imperfect.

To test this idea, in Dzindolet et al. (2003), three levels of feedbacks were presented to users (no feedback, cumulative feedback and continuous feedback). In Dzindolet et al. (2003), feedback is a bar graph which displayed the number of error made by human operators and the automatic aid during the trial. Dee and Valestine (2007) noted that automatic surveillance systems make four types of decisions:

1- False positive: the negative case that incorrectly identified as positive, i.e. detecting a wrong target.

2- True positive: positive cases that were correctly identified, i.e. detecting the right target.

3- False negative: positive cases incorrectly rejected, i.e. missing a right target

4. True negative: negative cases that were correctly identified as such, i.e. ignoring the wrong targets.

In the context of automatic surveillance system, if users know why the system has detected the target, then it might affect their trust and increase their performance. As noted in Dee and Velastin (2007), a threshold which defines the sensitivity of the system is set for every automatic surveillance system. If detection has been conducted through a low threshold the systems' confidence in that detection is less than detection that is conducted through a standard threshold.

2.5.5 Reliance

As described in the Riley (1996), a number of factors which directly influence the level of reliance of operator on automatic system are: confidence, fatigue, trust, state learning, and risk. In the context of surveillance, a CCTV operator should believe in his/her own judgement and do not completely rely on the automatic system. However if the task is tedious then the operator will have no other choice than to rely on the automatic system at same point. A complex automatic system which is hard to learn also prevents operators to rely on it; one reason is that when the system is complex it is more difficult to generate a mental model of the system and therefore the operator is effectively operating a system that he/she does not understand.

In a study conducted by Jamieson (2007), various features of the interface which affect operators' trust and reliance in an automated Combat Identification system have been investigated. This study suggests that usability is another aspect that can affect reliance and trust of operators on the automatic system.

A successful design of automatic surveillance system should encourage operators to rely on them without losing trust in their own judgment. It should also be easy to use and decreases the operators' workload.

2.6 Workload Assessment techniques

Workload is a cognitive concept which cannot be directly evaluated or observed. It should be inferred from a number of factors which describes its multifaceted nature. Methods of workload assessment and some examples of applying them for assessing workload of human operators while using various automatic systems such as autopilots, adaptive cruise controls and CCTV cameras have been reviewed in this section.

2.6.1 Primary and secondary task performance measures

Measuring primary and secondary task performance technique is an objective method which measures the performance of users while interacting with the automatic system. Participants have to perform two tasks; primary and secondary tasks. Primary task is the main task that the participant has to perform. However, as it is shown in Figure 2.4 primary task might not use all of the resource allocated to it and leave the operator with spare capacity. Hence if a user has an acceptable performance, this does not directly imply that the task demands are low. To deal with this issue, secondary task is introduced to the experiment. Participants are asked to attend to secondary tasks after completing the primary task. Measuring the performance of the secondary task will show the amount of spare capacity of mental resources.

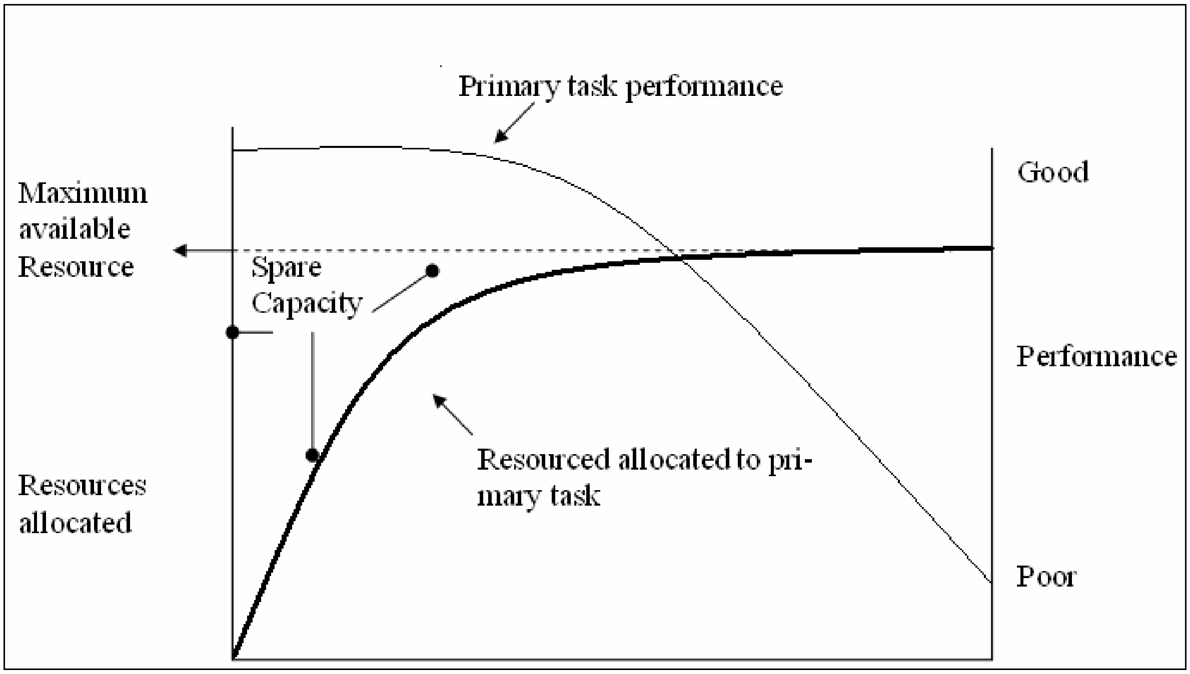


Figure 2-4: Resources allocated and demanded by primary and secondary tasks

Measuring primary and secondary task performance as a means for assessing workload has been widely used in literature. Park et al., (2007) have applied this technique to determine the performance of an operator while monitoring a multi display computer. The primary task in this experiment was typing while watching the main monitor and its secondary task was clicking on the targets that appear on the sub-monitor randomly. Both primary and secondary tasks in this experiment were realistic tasks in actual control rooms. Advantages of using secondary task in addition to primary task as reported in Sauer (2000) are that they are more sensitive and have higher face validity than other workload assessment techniques.

Secondary tasks should compete for the same resources as primary task to increase the demands for primary task. Sauer (2000) claimed that in most of the experiments available in the literature secondary tasks do not have an ecological validity. Hence the choice of appropriate secondary task is important to gather sound and valid estimation of the level of workload.

Young and Stanton (2004) assessed the workload of expert drivers while using an adaptive cruise control through primary and secondary task performance technique. The primary task in the experiment was to drive on a simple straight road for 10 minutes using the automated aid as well and a visual-spatial task was set for the secondary task. The results of the secondary task were recorded as a total of spare attentional resource available to drivers while using the automatic aid.

2.6.2 Physiological measures

Physiological measure is an objective technique to measure the users' physiological variations while interacting with the automatic systems. It is assumed that when the mental workload increases, operator's arousal will lead to changes in physiological measures such as cardiac activity, event-related brain potentials (ERP), eye blink activity and changes in the pupil diameter.

Physiological measures are the least common techniques since they are physically obtrusive. For example, various equipments have to be attached to participants to measure physiological variations. Moreover, their reliability depends on the technology and they are also expensive. However they have been used in cases where accurate and objective measurement of the MWL was required (Lee and Liu, 2003).

2.6.3 Subjective-rating techniques

As stated in Megaw (2005), subjective rating techniques are the ergonomist's favourite class of techniques. In subjective-rating techniques, participants will inform the analyst about their own workload; common methods are NASA Task Load Index (TLX), SWAT and Workload Profile.

Subjective Workload Assessment technique(SWAT) is a multidimensional subjective technique with three levels, time load, mental effort load and stress load, according to Stanton et al., (2005), this technique is quick and cheap but is has low sensitivity to mental workload. Workload profile is also a subjective multi dimensional method, it consists of eight mental workload dimension and aims to assess the demands upon the task through these dimensions, participants have to rate the dimension on a 0-1 scale. It is a fast and low cost technique which gives detailed information about various dimensions. However not so many studied have used this technique since it is new method and there is limited validation evidence (Stanton et al., (2005).

NASA-TLX is a multi-dimensional rating scale proposed by Hart and Staveland (1988). It provides information about the magnitude and sources of six workload-related factors to estimate the workload. Three of the dimensions reflect the demands the task puts on the operators (mental, physical and temporal) and the other three stress the interaction of operator with the task (performance, effort and frustration). It is considered to be the most effective method for assessing perceived workload (Warm et al., 1997).

Harris et al., (2008) conducted an experiment to investigate the effectiveness of geographical layout of CCTV screens on the performance and workload of the operators. These techniques are usually combined to provide the analyst with the best of both worlds. For example, Lee and Liu (2003) used both physiological and NASA TLX, Young and Stanton (2004) applied both primary and secondary task measures and NASA TLX to measure the drivers' mental workload.

3 Methods

The purpose of this project was to explore the effects of automation on operators' workload and performance. An experiment was designed and conducted to examine various features of automatic surveillance systems on operators' performance and workload. The experimental automatic surveillance system was mocked up through the rapid prototyping technique of "Wizard of Oz". In this technique, the tasks of the automatic systems are performed by a human "wizard".

The experimental system consisted of a set of video footages which were developed so that they resemble an automatic surveillance system. The video footage used for the experiment was recorded at Beeston railway station with an actor playing as the target. The raw video footage was then edited with Adobe Premier ™ 6.0 to simulate the automatic detection and tracking. Six different modes were defined for the experimental automatic system. Two accuracy groups were investigated and for each group raw video footage were recorded, these modes did not alarm users in any was, in other words in the baseline mode the operator has to find and track the target manually, in this report it is referred to as baseline mode. Detection and tracking was shown by drawing a box around the target. Furthermore to inform users about the systems' confidence in its choice, the boxes were colour coded where red represented high priority and orange represented low priority.

Assessment of workload was conducted by measuring secondary task performance and the NASA-TLX scales. The primary task was to detect and track the target and the secondary task was to count the number of people who were carrying a backpack.

This chapter explains the methodological strategy used to develop and implement the experiment and the techniques used for analysing the results.

Section 3.1 and 3.2 describes the procedure applied for mocking up the experimental automatic system. In section 3.1 the characteristics of the experimental automatic surveillance systems tested in this project is noted. Section 3.2 described the video shooting and editing required for the mock up. Section 3.3 descries the data gathering methods within the experiment, in section 3.4 details of participants is noted. Section 3.5 describes the apparatus used in the experiment. In section 3.6 the design of the experiment is explained, section 3.6 defines the method used to measure the variables of experiment and finally in section 3.7 the procedure of experiment is noted.

3.1 Experimental Automatic System's characteristics

3.1.1 Functionality

Functionality of the Experimental Automatic system in this project was limited to two tasks:

1- Detecting a target; in this report will be referred to as spotting

2- Tracking the target

These two tasks were separated to investigate operators' workload and performance in either of the cases to conclude which one is more complex for the human operator. The spotting mode of the Experimental Automatic system alert operator by drawing a box around the target when the target was first seen in the screen and the tracking mode alert the operator by drawing a box around the target when the target was leaving the screen. This is because tracking is generally to predict target's next location in the whole surveillance area and it is possible to do so if the operator knows where the target was last seen on the screen. Therefore the screen was divided into four zones and participants were asked to say the last zone in which the target was seen.

3.1.2 Confidence Information

The Experimental Automatic surveillance system of this project gives users information about how confidence the system's choice is. The choices made by the system are due to the level of systems sensitivity, in other words the events' characteristics might reach to the threshold which is considered to be suspicious. Setting an appropriate threshold to encounter environmental variability is challenging as mentioned in previous chapter. Figure 3.1 show various levels of threshold and the characteristics of the choices in each group.

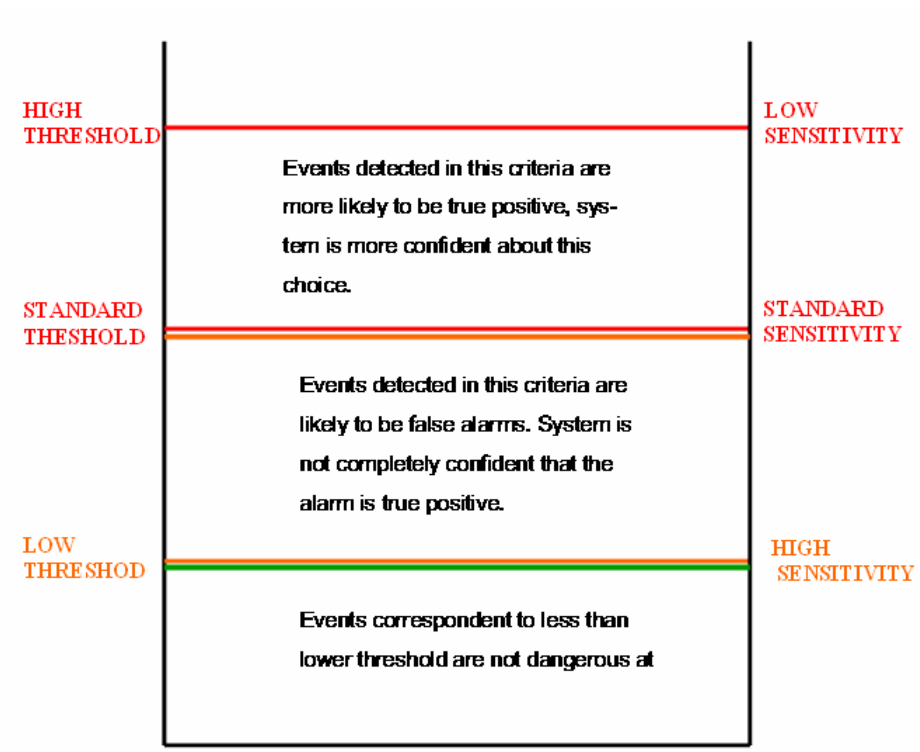


Figure 3-1: Confidence in the alarms in different threshold levels

All of the cases considered to be suspicious are tested against both of the thresholds. If it is only detected through the lower threshold, then the system's confidence about that choice to be true positive is low. However if the detection reaches the level of standard threshold, the choice is more likely to be true positive. This information is referred to as "confidence information", which is different from users' confidence in the system. It is identical to the case when a CCTV operator guess about an event to be suspicious, but some of his guesses are more likely to be true than the others. In this project the automated system inform users about the likelihood of that guess (alarm) to be true.

There are three levels of confidence information defined in this project: 1- no confidence information, 2- unreliable confidence information and 3- reliable confidence information. In the "No confidence information" level, alarms are generated by the system but system will not give users any clue about the likelihood of that alarm to be true positive. In the other two groups the system again generate alarms but differentiates its strong guesses and weak guesses, in more technical terms it differentiates the cases detected by standard threshold and cases detected by lower threshold with different colours to inform operators about its confidence in those alarms.

However, in the "unreliable confidence information", this differentiation is not accurate and there are cases in which the true positive case (the correct target) might be labelled as being detected with lower threshold. The reason for considering "unreliable confidence information" was since the experimental automatic system in this project was a wizard of oz mock up, the feasibility of developing such system was not clear. Hence if the system cannot clearly differentiate between the threshold levels, does this technique still improve performance and lower the demand?

In "reliable confidence information" mode the differentiation between the cases detected by standard threshold and cases detected by low threshold is correct, and system has higher confident in the cases detected with standard threshold whereas it is less confident in cases detected by lower threshold.

3.1.3 Accuracy

In order to study the effect of accuracy levels on operators' workload and performance, two accuracy groups have been chosen, 66% and 75%. Note must be taken that the definition of accuracy was simplified in the current project. Only ratio of the true positive to total number of detection was calculated and the inaccurate cases were limited to false alarms. According to Bliss (2003), the cost of alerting a false alarm is greater than misses. False alarms will lead to confusion and operators will completely lose trust in the system (Dixon, and Wickens, 2006)

The first level of accuracy, 66%, is the mean accuracy of the target recognition and tracking systems derived from the literature. For the tracking task, two recent automatic surveillance systems were chosen with accuracies of 85%(Muller-Schneiders et al., 2005) and 76% (Black, et al., 2005). The mean accuracy of the two systems (mean=80%) was set for the tracking function of the automatic surveillance system in this project. For the target recognition, a system developed by Shan et al. (2007) was chosen with the accuracy of 57%. Overall the automatic surveillance system of this experiment has the accuracy of 66% which is the mean of the accuracy of both functions (mean of 80% and 57%). In order to investigate the workload and performance of operators in a more accurate system, a second level of accuracy (75%) was applied.

In the 66% accuracy level, from the three alarms generated by the system, two of them are correct and the third is false alarm (2/3=0.66). In the 75% accuracy group, from the four alarms generated by the system three of them are correct and the forth is false alarm (3/4=0.75). This also affects the design of baseline video clips. In the 66% accuracy baseline the target appeared two times and in the 75% accuracy baseline the target appeared three times to ensure similar duration of target appearance within baselines and their correspondent experimental automated system.

3.1.4 Visualisation

In order to differentiate various modes of "confidence information" as discussed in the previous section, a colour coding system was used. The colour red is been conventionally used to attract attention and inform of dangerous situations. Furthermore based on the traffic light stereotype, orange is also associated with danger but is not perceived to be as critical as red.

In the "no confidence information" case, since the system gives no information about its confidence in the choice of alarms, the entire detected cases are considered to be equally dangerous and hence they are all shown with red. However in reliable and unreliable confidence information modes, the system needs to differentiate between the highly confident cases and cases which are less likely to be true positive. Red was used when the system was confident about its choice (it was detected with the standard threshold). And orange was used when the target was detected through the lower threshold. Note must be taken that in the "unreliable confidence information", this differentiation is not necessarily correct, there are cases that the true positive case which have been detected through an standard threshold was shown with orange.

The coloured box appeared around the target for 500 ms which is a sufficient duration to be noticed (Wickens et al., 2004). Eye movement which is necessary to search the visual field is generally quite fast (less than 50 ms). Moreover there are various factors which have an impact on visual search such as eye contrast sensitivity, how experienced the user is and how focused he/she is on the scene (Wickens et al., 2004)

In sum, the Experimental Automatic system with two types of functionality (spot and track) has to be tested in two different accuracy levels (66% and 75%). In each of these conditions the three levels of confidence information were applied. In each of the accuracy levels, six different modes were mocked up for the experiment:

1- No confidence Spot (NS)

2- No confidence Track (NT)

3- Unreliable confidence Spot (US)

4- Unreliable confidence Track (UT)

5- Reliable confidence Spot (RS)

6- Reliable confidence Track (RT)

Furthermore a baseline was required to enable the comparison of automatic and non automatic systems; in each accuracy group, two video clips were kept as raw video footage of the target. The only reason of having two video clips was to have more accurate baseline results. This means that 16 video clips had to be prepared for the experiment.

3.2 Video and Audio

Beeston railway station was chosen for recording the video with two platforms. It is not a very busy station. According to the information provided by the Office of Rail Regulations (ORR), (www.rail-reg.gov) in 2006-2007 the number of annual passengers originating and ending Beeston railway station was 368,000. There is one train every 20 minutes and there are times when the station is empty. The busiest times of the day are between 7.30- 8.45 am and 17.00-18.45 pm, which was chosen as the time for the filming. An actor was recruited and the description of his physical characterises have been recorded to be used in the experiment.

3.2.1 Audio Description

20 seconds audio description of the actor's physical characteristics was recorded, the content of the audio description was:

"The suspect is a white male about 25 years old, 180 cm tall and 90 kg weight. He has dark brown hair and is alone, without bag, wearing a brown T-shirt and short trousers."

3.2.2 Video shooting

In order to control the conditions of experiment and reduce the number of variable, one filming spot was used for all of the sixteen video clips. The spot should provide a reasonably good view of both of the platforms and all entrances (Figure 3.2) and the camera should be unobtrusive (Figure3.3).

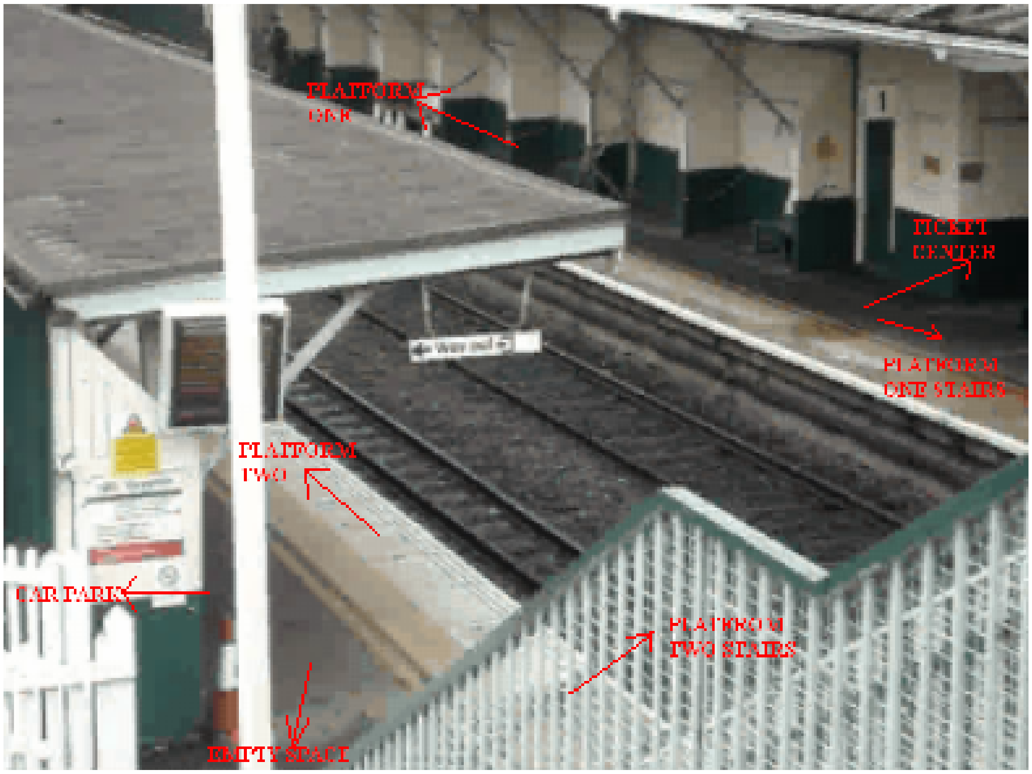


Figure 3-2: Field of view

Figure 3-3: Position of the camera

70 minutes of video was recorded using a Sony DSC-W55 Camera, 7.2 Mega pixels Cyber shot camera. These video clips were then edited to meet the conditions of the experiment.

3.2.3 Video Editing

Adobe Premier ™ 6.0 was used to edit the video. The main purpose of the editing was to prepare the video clips as if there is an automated system installed on them. Each video clip represents one mode of the experimental automatic system.

The video clips are each two minutes long and they have been made from the total of 70 minutes video footage. The two minutes was chosen due to the nature of reactive monitoring and the fact that CCTV operators have to take a break every 20-30 minutes based on health and safety regulation (Wallace and Diffley, 1998). Since 8 trials had to be performed, if trials were longer than two minutes the workload would increase due to fatigue and other extraneous variables.

In order to make the clips identical and comparable, various factors have been considered. Each clip had to give participants similar views of the target, the field of view of the camera had better resolution over platform two and hence the clips were edited in a way that the target appeared in both of the platforms on every video clip. Furthermore, the interval of appearance of the target in the video clips had to be approximately similar (mean=36 seconds). Another factor to consider was the duration of the presence of the actor which had to be approximately similar in every clip (mean=24 seconds). Table 3.1 shows the outline of the video clips in the 66% accuracy level.

	First appearance of the target		Second appearance of the target			
Clip Number	Platform	Time of appearance (in 2minutes video clip)	Platform	Time of appearance (in 2 minutes video clip)	Interval between the two appear-ances (Seconds)	Total Duration of appearance (Seconds)
1	Two	0.35	One	1.10	35	26
2	One	0.50	Two	1.25	35	28
3	One	0.10	Two	0.55	45	22
4	Two	0.20	One	0.55	35	22
5	One	1.00	Two	1.35	35	26
6	Two	0.30	One	1.05	35	20
7	One	0.45	Two	1.25	30	27
8	Two	1.15	One	1.50	35	20
MEAN					36	24

Table 3-1: Details of the video clips in 66% group

Similarly in the 75% accuracy level, the mean duration of appearance of the target was 28 seconds and the interval between the three appearances was 31 seconds. This was deliberately chosen to be similar to the 66% accuracy level to keep the variables of the two experiment limited to the accuracy of the systems. The target appeared three times on these video clips. Table 3.2 shows the detail of duration and interval of presence as well as the platform number in which the actor appeared in the eight video clips.

	First appearance of the target		Second appearance of the target		Third appearance of the target			
Clip Number	Platform	Time of appearance (in 2minutes video clip)	Platform	Time of appearance (in 2minutes video clip)	Platform	Time of appearance (in 2minutes video clip)	Interval (Seconds)	Duration (Seconds)
1	Two	0.35	One	1.10	Two	1.35	30	32
2	One	0.15	One	0.50	Two	1. 25	35	32
3	One	0.10	Two	0.55	One	1.30	40	25
4	Two	0.20	One	0.55	Two	1.35	37.5	27
5	Two	0.25	One	1.00	Two	1.35	35	29
6	Two	0.30	One	1.05	Two	1.40	35	26
7	One	0.45	Two	1.25	One	1.40	32.5	30
8	One	0.40	Two	1.25	One	1.50	40	25
MEAN							31.12	28.25

Table 3-2: details of video clips in 75% group

In order to prepare the video clips as shown in table 3.1 and 3.2, the 70 minutes video file was imported in a time line in the editor page of Adobe Premier ™ 6.0. Required seconds based on tables 3.1 and 3.2 were selected with the 'RAZOR' tool on the time line window of the Adobe Premier software and created a new video track. These were finally exported to the trimmed video file with AVI format so that it could be viewed by video players.

To edit a video clip for the spotting task, an image of the scene was captured when the target appeared in the screen, a coloured box was drawn around the target with Microsoft ® Paint (version 5.1) and placed back in the timeline of the clip with the duration of 500 ms. The same procedure was applied for the track function except in this case the image was captured when the target was leaving the screen and also apart from the coloured box around the target the zone number in which the target was last seen in was also printed inside the box.

Figure 3.4 and 3.5 show an example of spot and track in the automated prototype.

Figure 3-4: Example of Spot

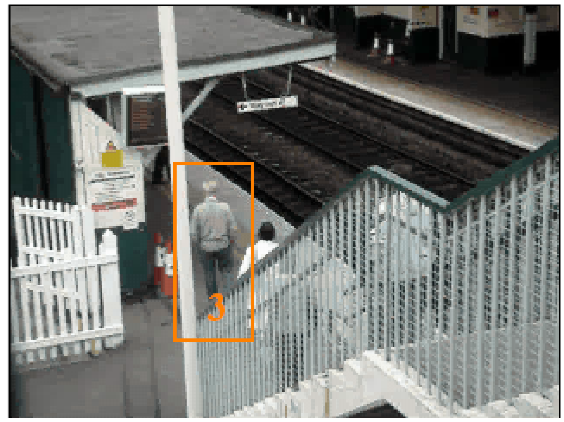


Figure 3-5: Example of track

To give detections various modes of "confidence information" a colour coding system was applied. Tables 3.3 and 3.4 show the outline of detected cases in both 66% and 75% accuracy levels and the colour codes used for each of the detections. In the 66% accuracy level, three alarms were triggered where two were correct and one was false alarm. In other words, two of the alarms are true positives and one is false positive. In "No confidence Spot" (NS) and "No confidence Track" (NT), since there was no information about the confidence of the detection, everything was considered equally dangerous (all red). In the "Unreliable confidence Spot" (US) and "Unreliable confidence Track" (UT), the system gave information about its confidence for each of the choices but this information was unreliable. Hence there were cases when the true positive was detected with an orange box. Finally, in "Reliable confidence Spot" (HS) and "Reliable confidence Track" (HT), the system gave reliable information about the confidence of system in its detection where the ones detected with the lower threshold were coded as less dangerous with colour orange and the ones detected with the standard threshold were coded with colour red.

Mode	First Alarm	Second Alarm	Third Alarm
No confidence Spot	True positive	False Positive	True Positive
No confidence Track	False Positive	True Positive	True Positive
Unreliable confidence Spot	False Positive	True Positive	True Positive
Unreliable confidence Track	True Positive	True Positive	False Positive
Reliable confidence Spot	False Positive	True Positive	True Positive
Reliable confidence Track	False Positive	True Positive	True Positive

Table 3-3: Details of alarms in 66% accuracy level

75% accuracy group followed the exact outline except that in this group the system reported four alarms, three of which were correct and one was false alarm.

Mode	First Alarm	Second Alarm	Third Alarm	Fourth Alarm
No confidence Spot	True positive	False Positive	True Positive	True Positive
No confidence Track	False Positive	True Positive	True Positive	True Positive
Unreliable confidence Spot	False Positive	True Positive	True Positive	True Positive
Unreliable confidence Track	True Positive	True Positive	False Positive	True Positive
Reliable confidence Spot	False Positive	True Positive	True Positive	True Positive
Reliable confidence Track	True Positive	False Positive	True Positive	True Positive

Table 3-4: Details of alarms in 75% accuracy level

3.3 Data gathering Methods

3.3.1 Interview

The experiment was conducted in the laboratory, but it was aimed to simulate real life situation of CCTV control rooms. In addition to the information collected in Chapter two, an informal interview was conducted with the crime prevention officer in University of Nottingham. This was a semi-structured interview with a set of open ended questions.

1- What are the main responsibilities of security staff?
2- How many cameras are installed in the University?
3- How they use CCTV camera?
4- How often they monitor proactively?
5- What do they mean by suspicious behaviour?

The interview revealed that although there are 170 CCTV cameras installed in all campuses of university of Nottingham the monitoring is mostly reactive. Moreover the archives are held for up to 21 days which is used for forensic analysis. The only proactive monitoring is a daily check of cameras to make sure they are running and occasionally checking safety critical locations. The crime prevention officer also noted that majority of cases which lead to the use of CCTV cameras are when there is a report about suspicious event through radio channels.

In the experiment designed in the present project, the researcher has used pre-recorded audio description of an actor as the source of information. The participants were asked to find the target after they have listened to its description, hence the surveillance was reactive.

3.3.2 Questionnaires

- NASA-TLX

Assessing the workload was done using the 'Raw' version of the NASA-TLX, which gives equal weight to all dimensions (Byers et al., 1989). There are six sub dimensions in this questionnaire: mental demand, temporal demand, physical demand, performance, frustration and effort. The scale of each of the dimensions ranged from low (0) to high (100). The overall workload score was calculated as a mean of all six dimensions. The overall workload score is referred to as RTLX.

- Reliance Questionnaire

In order to assess influential factors on operators' reliance on automation a short questionnaire was designed .This questionnaire consisted of six questions which participants had to rate from low (0) to high (100). Four of the questions were related to factors proved to be influential in a study by Riley (1996) which are: trust, state learning, confidence, risk. Riley's automation reliance factors and the questions used to subjectively measure them are listed below:

1- Trust→ how much trust you have in the system.

2- State learning→how did you understand the information presented to you?

3- Confidence→how can you spot or track better than the system?

4- Risk→ how misleading was the system to use?

Similar to the experiment conducted by Jamieson (2007), effectiveness was used as the fifth factor. It was investigated through the two other questions in the reliance questionnaire of this project. The complete questionnaire has been presented in Appendix II.

1- How easy was the system to use?

2- How helpful did you think the automatic system was?

Both of these questions aimed to reflect the subjective opinion of users regarding how they achieved the goal (spotting and tracking) with the aid of Experimental Automatic system.

3.4 Participants

24 University students volunteered to participate in this study.11 male and 13 female with an average age of 22.7 years old. They had been divided into groups of 12 and each group conducted the experiment with one of the accuracy levels (66% or 75%). They did not have any prior experience with monitoring CCTV video footages. Ethical guidelines of the University of Nottingham were strictly followed throughout the experiment.

3.5 Apparatus

The mocked up Experimental Automatic system was developed by Adobe Premier™ 6.0, it was used for editing the video footage which has been recorded by Sony DSC-W55 Camera, 7.2 Mega pixels Cyber shot camera on a tripod. The audio description was recorded by a digital MP3 player.

The edited video clips were shown to participants in a Sony VAIO laptop, PCG-V505EX, using a JetAudio player (Windows edition) .The audio description was played in a Sony MP3 player.

3.5.1 Other Considerations

1. Pre-experimental test

To ensure that participants do not know the target, 10 random images of people including the actors' image was shown to the participants. After participants confirmed that they did not know the target, the experiment was started.

2. Rules reminder sheet

To help participants constantly remember the conditions of the automatic modes each time, they were provided with a rules reminder sheet were the details of each of the automatic modes was briefly explained. Prior to each viewing of video clips the participants were told which mode will be running next so that they could review the details of that mode.

3. Zone numbers

The screen was divided into four zones for the tracking task to help participants know the zone numbers, a sheet was provided to participants with the zone numbers on it.

4. Result sheet

In order to keep track of the performance of participants a results sheet was designed. This sheet had four sections for keeping track of primary task performance and a column for the performance of secondary task performance. All of the documents explained in this chapter are in Appendix II.

3.6 Design

For each of the accuracy groups a within subject design was used with 12 participants viewing all eight video clips (2 baselines and 6 automated). To prevent the effect of learning and also participant fatigue, the automated modes were counterbalanced through a 6 by 6 Latin Squared table (Appendix I).

There are two independent variables: 1- the functionality of automated system with two levels, spot and track and 2- the confidence information level: no confidence information, unreliable confidence information and reliable confidence information which is combined into six modes of automatic systems. There was also a baseline mode which is raw video footage which does not generate and alarm and operator have to detect and track the target manually.

1. No confidence Spot (NS)
2. No confidence Track (NT)
3. Unreliable confidence Spot (US)
4. Unreliable confidence Track (UT)
5. Reliable confidence Spot (RS)
6. Reliable confidence Track (RT)
7. Baseline

Dependent variables were workload RTLX score obtained from NASA-TLX, secondary task performance and reliance factors.

The six modes have been tested through a within subject design of the 12 participants in each of the accuracy groups. The hypotheses investigated were:

H1: The main effect of functionality on the dependent variables.

H2: The main effect of “confidence information” level on the dependent variables.

H3: The main effect of interaction of functionality and “confidence information” level on the dependent variables.

The two accuracy groups were tested with a between subject design of the 24 participants. The hypotheses investigated in this test are:

H1: there is a significant difference between the dependent variables in the automated modes and the dependent variables in the baseline.

H2: there is a significant difference between the dependent variables of the two accuracy levels.

3.7 Experiments’ Measures

Workload was assessed through NASA-TLX Raw version, where all of the six dimensions are assumed to have equal weights, the total workload score is the mean of the values for the six dimensions and it is referred to as RTLX score (Byers et al., 1989).

The performance of secondary task was also measured as the ratio of the number of backpacks counted by participants in each mode to the actual number of backpacks in each mode. The percentage obtained from this ratio considered to be the amount of spare attentional capacity in that mode.

The reliance questionnaire which focused on five factors also is rated from 0-100. For four of the questions which were correspondent to the factors proposed by Riley (1996), this percentage is recorded. For effectiveness which was correspondent to the two other questions, the average of the two was measured.

3.8 Procedure

The recruitment of participants was done using flyers around the University of Nottingham campuses. The participants were briefed about the experiment. After ensuring that the target was not known by the participants through the pre-experimental test, they listened to the audio description of the target. Since the mode of each video change after each trial a rules reminder sheet was provided together with the zone numbers.

Participants were allowed to listen to the description twice and they could take note of it. Baseline scenarios were presented first and were followed by NASA-TLX forms. Then participants viewed the six automated video clips in a counter balanced order, and were asked to fill in both NASA-TLX and Reliance questionnaire. Note must be taken that they were asked to find and track the suspect in all of the video clips as their primary task and also had to count the number of people with backpacks as their secondary task.

Finally they were debriefed about the experiment and filled a demographic questionnaire related to their gender and age and were rewarded 8 pounds as a gift for their participation in the experiment.

4 Results

This chapter reports the results obtained from the experiment. The results have been analysed statistically using SPSS 15.0. The statistical tests applied were within subject 2 by 3 repeated measures ANOVA, within subject paired sampled t-test and between subjects independent sampled t-test.

Section 4.1 compares the RTLX scores and secondary task performance scores for the baseline modes in both of the accuracy levels; section 4.2 reports the results of RTLX score, secondary task performance and reliance factors in all of the experimental automated conditions. Section 4.3 compares the automated modes and baselines in terms of their RTLX scores and secondary task performance score for both of the accuracy levels and finally section 4.4 reports the participants' comments about how they used the audio description to detect the target.

4.1 Experimental Baseline System

Baseline mode, which is a representative of a manual system, does not provide the operator with any automatic aid. Both of the accuracy levels (66% and 75%) have a baseline case. Since they are not automated, they are identical except for one thing. As noted in the previous chapter, in the 66% group the actor (target) appeared twice whereas in the 75% group the actor appeared three times. This was the only condition that was considered in the baseline scenarios.

The assumption is that the RTLX score and secondary task performance scores of the baseline modes in the two accuracy groups should be similar. The results obtained from NASA-TLX of the two baseline mode (Figure 4.1) show that the baseline mode with 75% accuracy is less demanding in comparison to the 66% accuracy level's baseline.

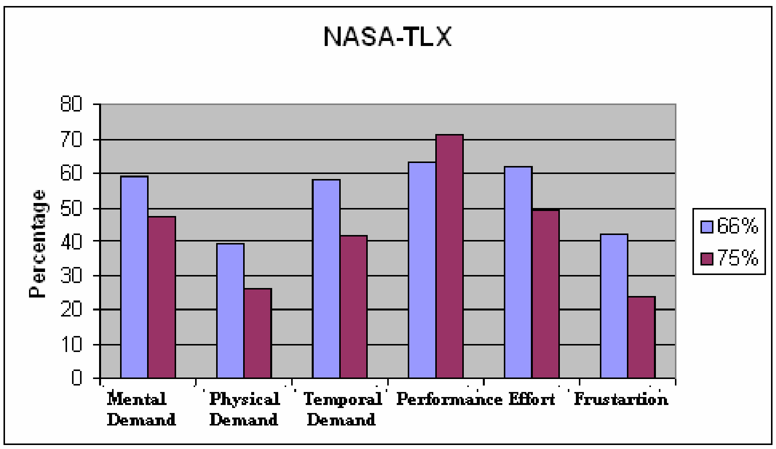


Figure 4-1: NASA-TLX dimensions for both baselines

The RTLX score was calculated in both of the cases and the results are presented in Figure 4.2.

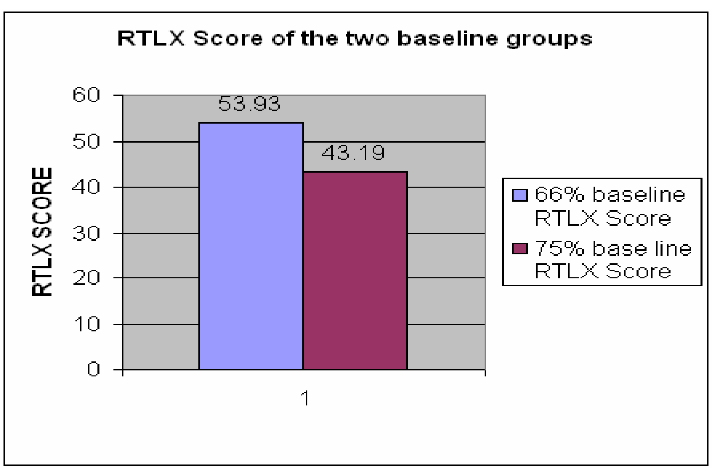


Figure 4-2: RTLX score of the baseline modes in 66% and 75% accuracy level

Furthermore, in order to obtain a value for the spare amount of attentional capacity, the percentage of secondary task performance was calculated. Figure 4.3 shows the percentage of secondary task performance in each of the baseline groups.

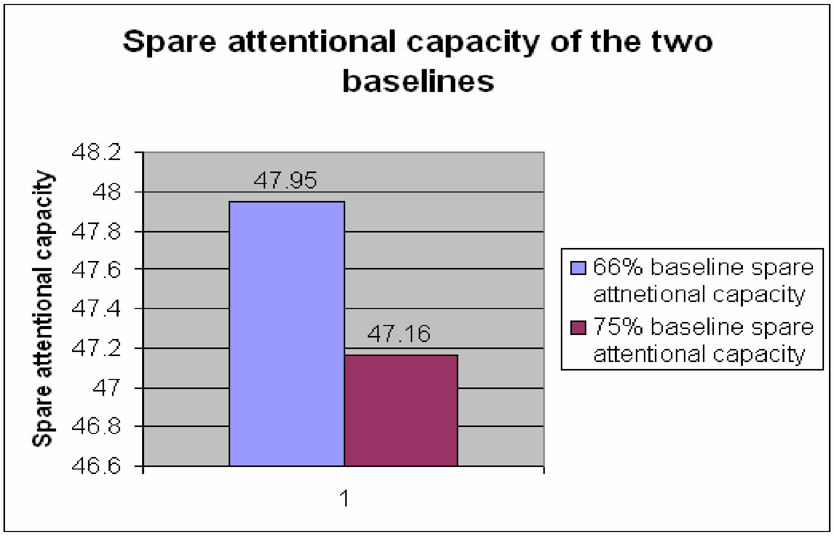


Figure 4-3: Spare attentional capacity of the baseline modes in 66% and 75% accuracy level

This graph shows that the secondary task performance of the baseline at the 66% level is 0.7 % higher than the secondary task performance of the 75% level. However, applying paired sampled t-test between secondary score of the two accuracy groups and the RTLX score of the two accuracy groups shows no significant difference ($P>0.05$).

4.2 Experimental Automated Systems

This section investigates the results obtained from both accuracy groups' automated modes. As mentioned in chapter three, there are six automated modes for each of the accuracy groups.

4.2.1 66% Accuracy Level

4.2.1.1 NASA-TLX

The Automatic clips in the 66% accuracy group have been tested using a within subject 2 by 3 repeated measures ANOVA. The hypotheses investigated in this project are:

H1: The main effect of functionality on the RTLX scores.

H2: The main effect of "confidence information" on the RTLX scores.

H3: The main effect of interaction of functionality and "confidence information" on the RTLX scores.

Before applying the ANOVA, the data have been tested for the normality and homogeneity assumptions. The cases which were not normal or were heterogeneous were transformed using BoxCox transformation. Figure 4.4 shows the RTLX score of the six automated modes in the 66% accuracy level.

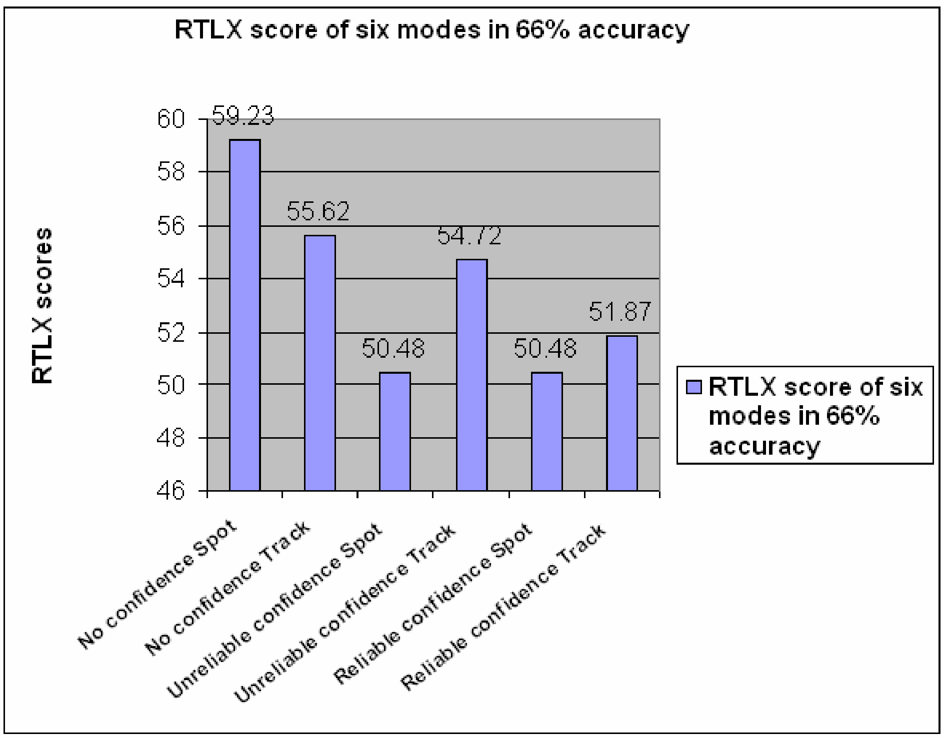


Figure 4-4: RTLX score of 66% accuracy group

Giving "confidence information" of the system had significant main effect on the RTLX scores, $F(2, 10) = 6.910$, $p<0.05$. Post hoc analysis among the three "confidence information" levels shows that there is a significant difference between "no confidence" level and "reliable confidence" level ($p<0.01$). Estimated marginal means of the RTLX scores shows that when the system gives reliable confidence information of its detection, the RTLX score have the least value whereas when there is no confidence, the RTLX score has the highest value (Figure 4.5).

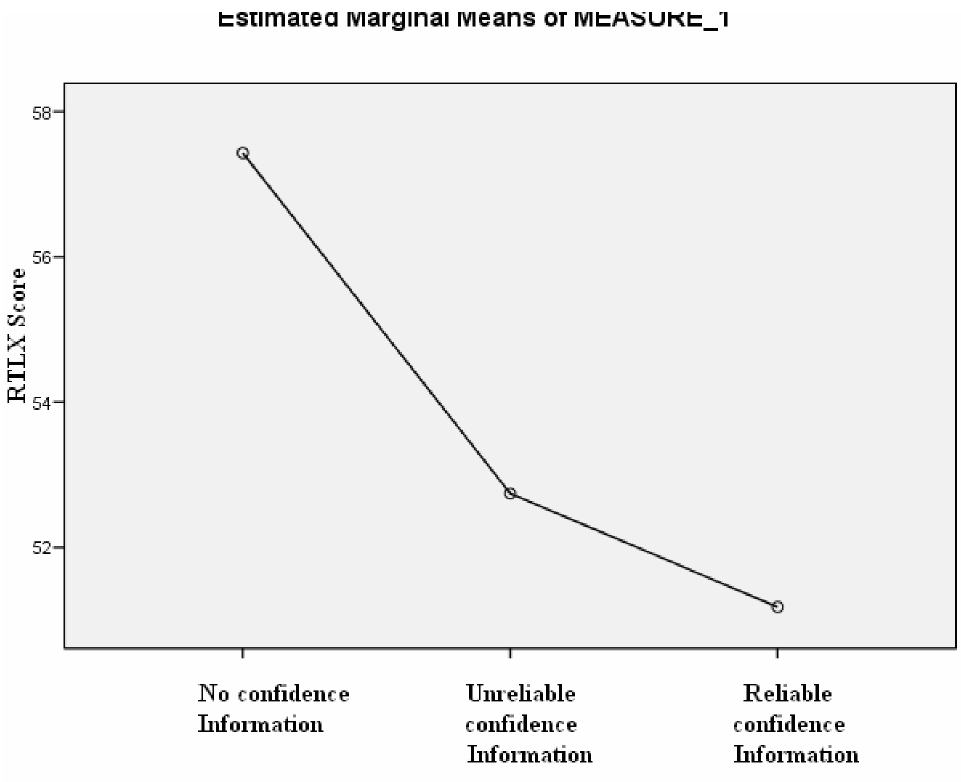


Figure 4-5: Marginal Mean of the three "confidence information" level on RTLX score of 66% accuracy group

Each of the six NASA-TLX dimensions (mental, physical, temporal, performance, frustration and efforts) has also been tested for the three hypotheses mentioned. Providing "confidence information" to operators had significant effect on mental demand, temporal and effort. Post hoc analysis of the three levels of providing "confidence information" showed that there is a significant ($P<0.05$) difference between "no confidence information" and "reliable confidence information". The three other NASA—TLX dimensions (performance, frustration and physical demand) were not significantly affected neither by providing various levels of "confidence information" nor by introducing different types of functionality.

4.2.1.2 Secondary Task Measures

A 2 by 3 repeated measures ANOVA test was applied to investigate the following hypotheses regarding the secondary task:

H1: The main effect of functionality on the secondary task performance.

H2: The main effect of the "confidence information" level on the secondary task performance.

H3: The main effect of the interaction of "confidence information" and reliability on secondary task performance.

Figure 4.6 shows the mean value of the secondary task performance in the 66% accuracy group.

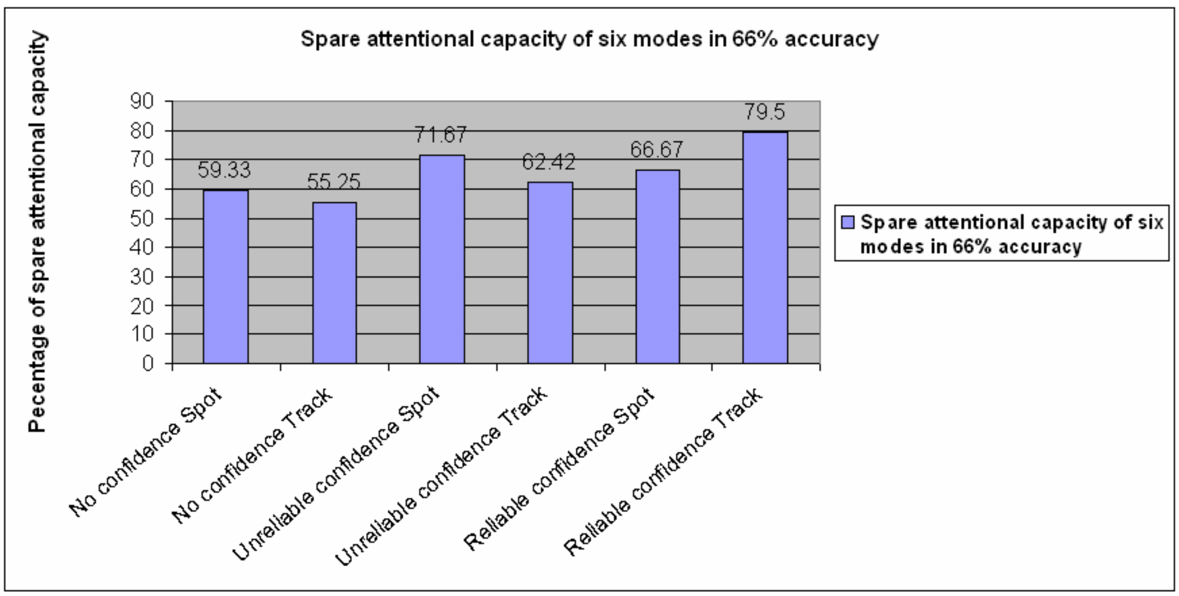


Figure 4-6: secondary task performance of 66% accuracy

Having a "confidence information" significantly affects the capacity of attentional resources, $F(2, 10) = 6.552$ and $P<0.05$. Post hoc analysis revealed that the difference between the "No confidence information" and "Unreliable confidence information" ($P<0.05$) as well as "No confidence information" and "Reliable confidence information" ($P<0.05$) are also significant whereas the difference between the "Unreliable confidence information" and "Reliable confidence information" is not significant (P>0.05). Based on the estimated marginal means, a system with "Reliable confidence information" has the highest performance of secondary task (Figure 4.7). No other significant effects were observed.

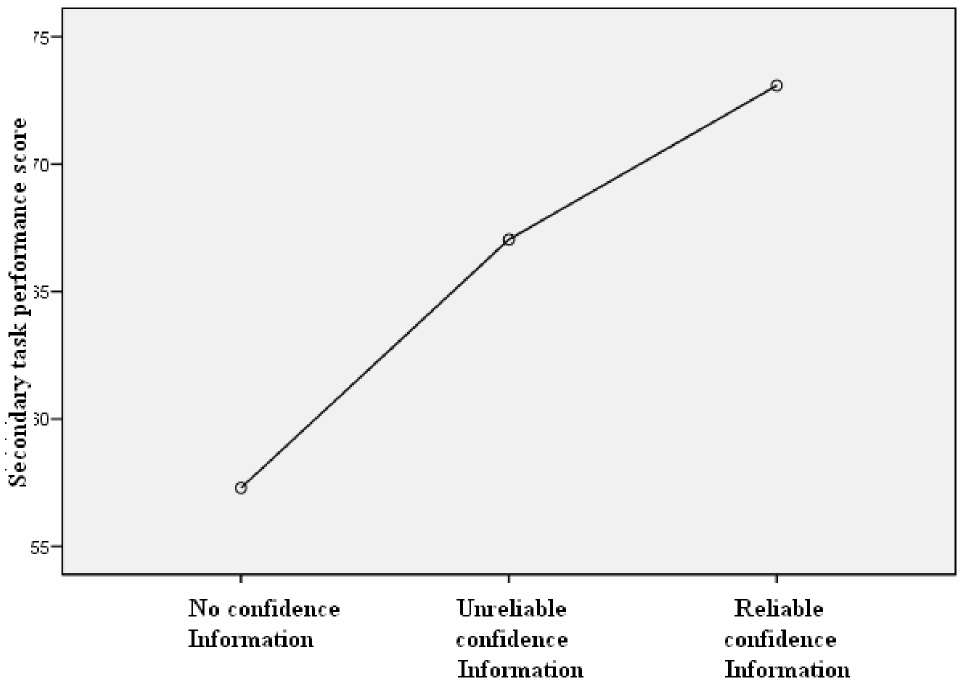


Figure 4-7: Marginal Mean of the three "confidence information" level on secondary performance score of 66% accuracy group

4.2.1.3 Reliance Factors

A series of 2 by 3 repeated measures ANOVA tests have been applied to investigate the following hypothesis related to the reliance factors:

H1: The main effect of functionality types on the five reliance factors.

H2: the main effect of "confidence information" levels on the five reliance factors.

H3: the main effect of the interaction of functionality types and "confidence information" levels on the five reliance factors.

As mentioned in chapter three, the Reliance questionnaire was designed to subjectively investigate five main factors associated with reliance of CCTV operators on automatic surveillance systems: trust, state of learning, effectiveness, confidence and risk. Each of these dimensions has been tested separately using a 2 by 3 repeated measures ANOVA.

None of the reliance factors has been significantly affected by the level of "confidence information" or functionality types. Table 4.1 shows the independent variables of functionality with two levels (Spot and Track) and levels of confidence information (no confidence, Unreliable confidence information and Reliable confidence information) which had the highest scores in each of the five reliance factors.

Reliance factor	Functionality level	Confidence information Level
Effectiveness	Track	No confidence information
Trust	Track	Reliable confidence information
Risk	Spot	Unreliable confidence information
Confidence	Track	Unreliable confidence information
State of Learning	Track	No confidence information

Table 4-1: highest scores for the five reliance factors in 66% accuracy

4.2.2 75% Accuracy Level

4.2.2.1 NASA-TLX

The Automatic clips in the 75% accuracy group have been tested using a within subject 2 by 3 repeated measures ANOVA. The hypotheses investigated in this section are:

H1: The main effect of functionality types on the RTLX scores.

H2: The main effect of various "confidence information" levels on the RTLX scores.

H3: The main effect of interaction of functionality types and levels of "confidence information" on the RTLX scores.

The RTLX of the 75% accuracy group is presented in Figure 4.8.

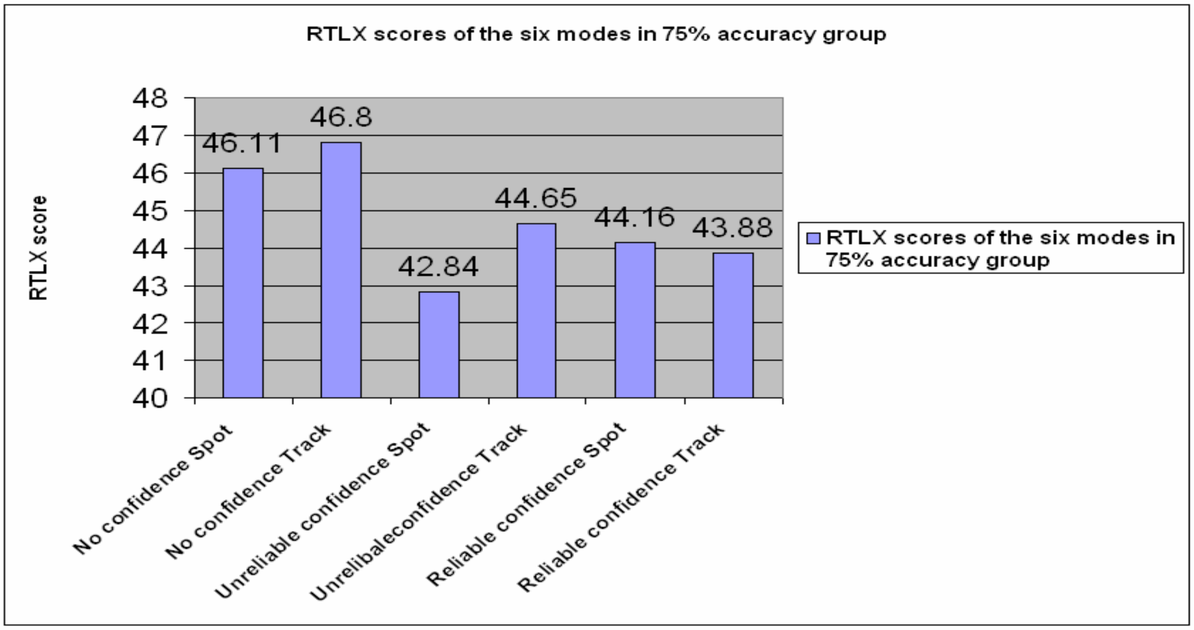


Figure 4-8: RTLX score of 75% accuracy group

Providing confidence information significantly affected the RTLX score, $F(2, 10) = 6.910$, $p<0.05$. Post hoc analysis showed that providing "unreliable confidence information" to users significantly decreases their RTLX score compared to the case where the aid gives "no confidence information" ($P<0.05$). Figure 4.9 shows the estimated marginal means of RTLX score in the three "confidence information" levels. No other significant effects were observed.

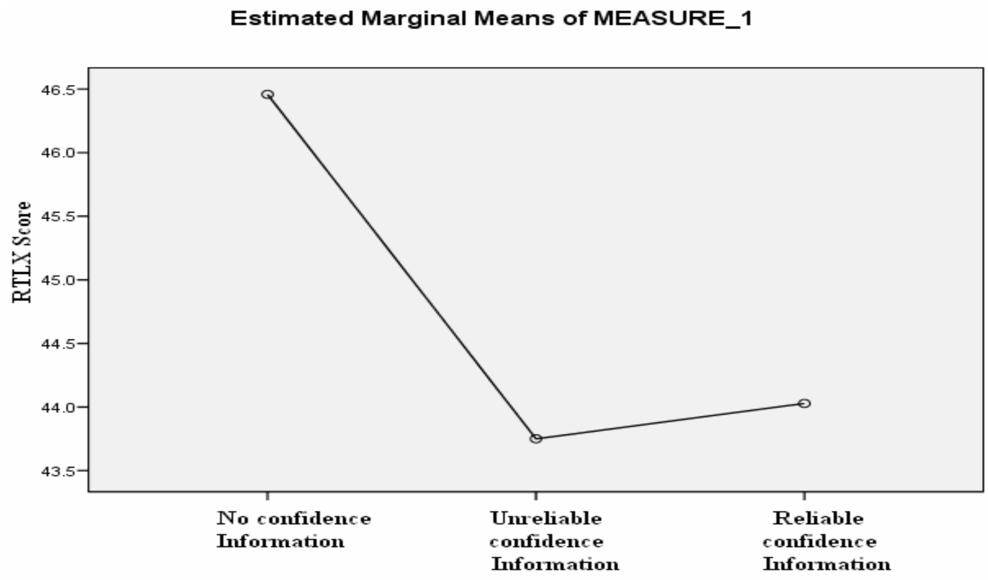


Figure 4-9: Marginal Mean of the three "confidence information" level on RTLX score of 75% accuracy group

The six dimensions of NASA-TLX (mental demand, physical demand, temporal demand, performance, frustration and effort) have also been tested to investigate the main effect of "confidence information" and functionality on each of the dimensions using 2 by 3 repeated measures ANOVA and again no significant effect was observed.

4.2.2.2Secondary Task Measures

Figure 4.10 shows the mean score for the secondary performance in the 75% accuracy group. A 2 by 3 repeated measures ANOVA test was used to investigate the following hypotheses:

H1: The main effect of functionality on the secondary performance scores.

H2: The main effect of "confidence information" level on the secondary performance scores.

H3: The main effect of interaction of functionality and "confidence information" level on the secondary performance scores.

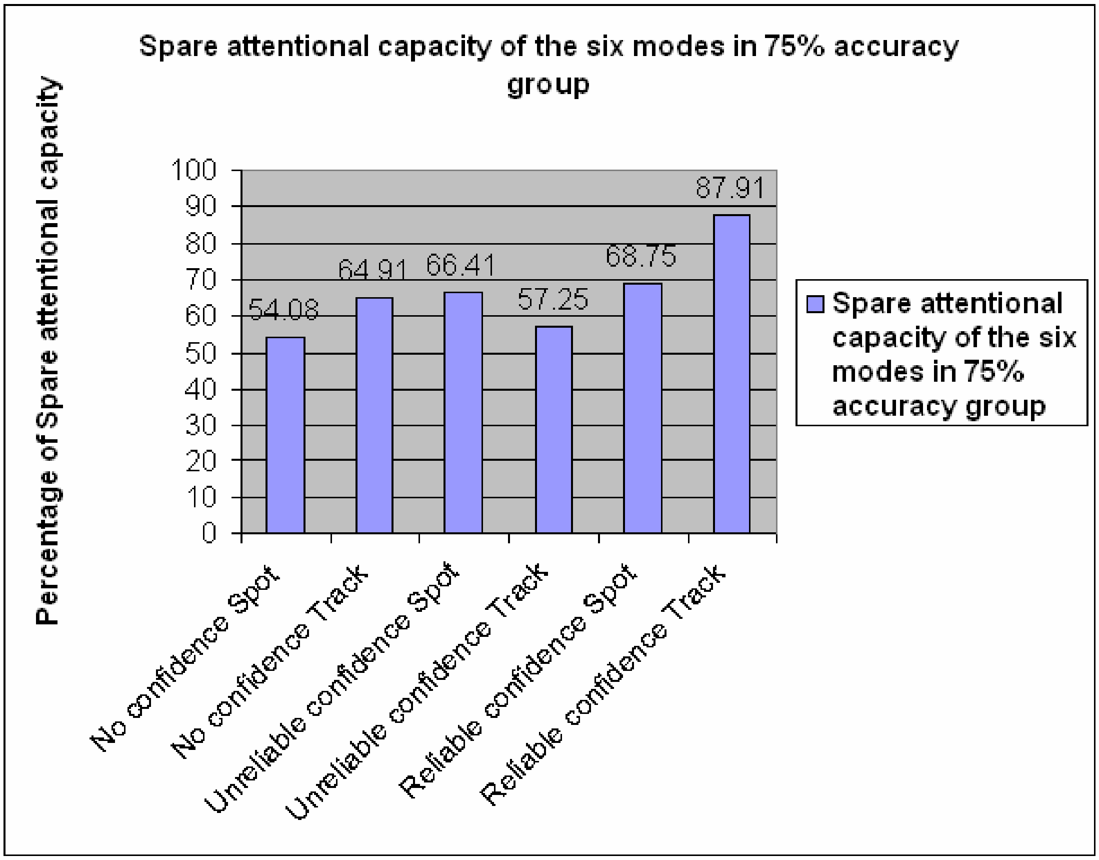


Figure 4-10: secondary task performance of 75% accuracy group

"Confidence information" levels significantly affect the secondary performance score, $F(2, 10) = 4.5111$ and $P<0.05$. Furthermore, the main effect of the interaction of "confidence information" levels and functionality is also significant $F(2, 10) = 6.276$, $P<0.05$. Post hoc analysis of the "confidence information" levels shows that there is a significant difference between "no confidence information" and "reliable confidence information" ($P<0.05$).

Figure 4.11 shows estimated marginal means of secondary performance in the three "confidence information" levels.

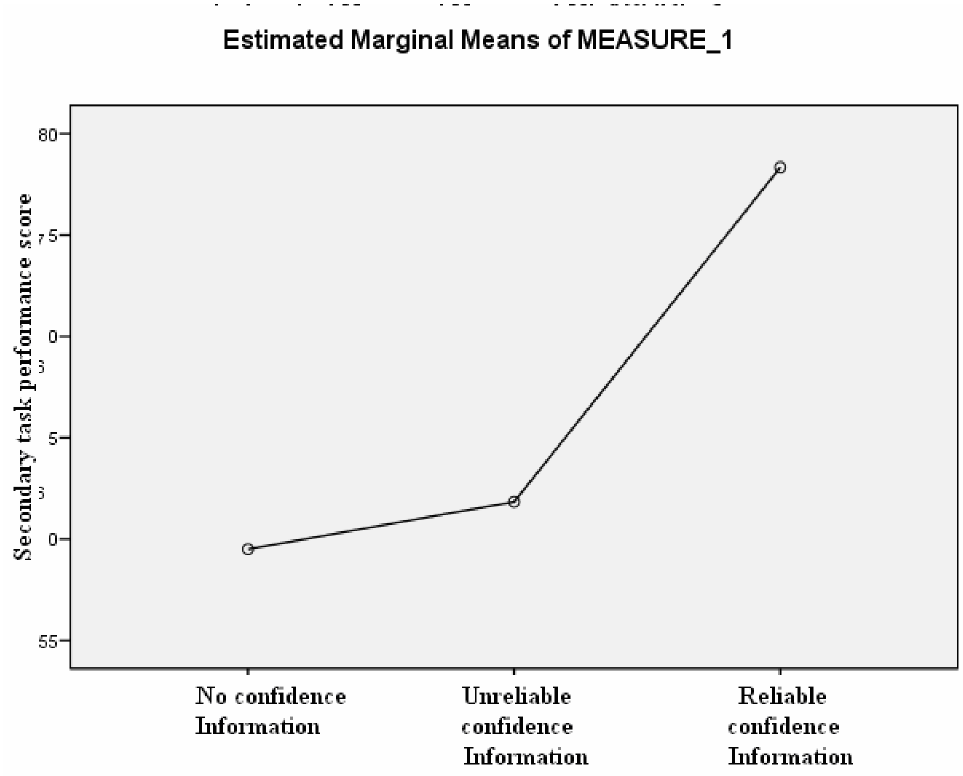


Figure 4-11: Marginal Mean of the three "confidence information" level on secondary performance score of 75% accuracy group

4.2.2.3 Reliance Factors

The five reliance factors obtained from the reliance questionnaire were tested by a 2 by 3 repeated measures ANOVA test in order to investigate the following hypotheses:

H1: The main effect of functionality on the five reliance factors.

H2: the main effect of "confidence information" levels on the five reliance factors.

H3: the main effect of the interaction of functionality and "confidence information" levels on the five reliance factors.

The main effect of "confidence information" level was found to be significant on two of the factors: risk and trust. No other significant effects were observed.

For risk: F (2, 10) =7.953, P<0.05, post hoc analysis of the three “confidence information” levels shows a significant difference between “no confidence information” level and “reliable confidence information” level (P<0.05) .Figure 4.12 shows the estimated marginal means of risk score in the three levels.

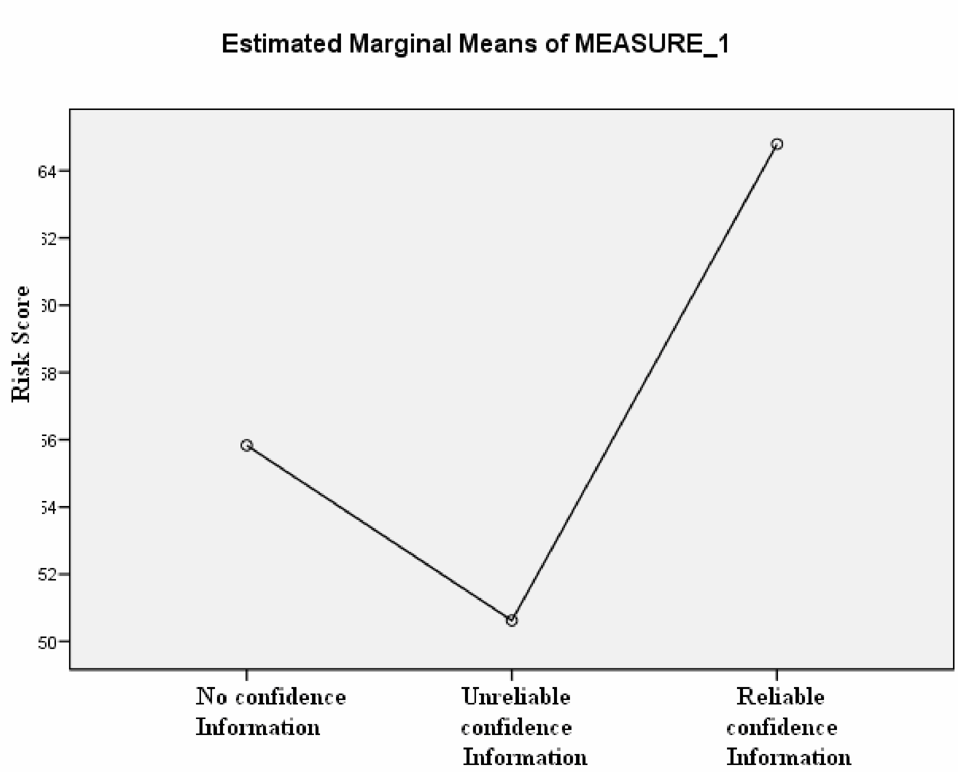


Figure 4-12: Marginal Mean of the three “confidence information” level on risk score of 75% accuracy group

For trust, F (2, 10) =7.816, P<0.05, post hoc analysis shows significant difference between “unreliable confidence information” and “reliable confidence information” levels (P<0.05). Figure 4.13 shows the estimated marginal means of trust score in the three levels.

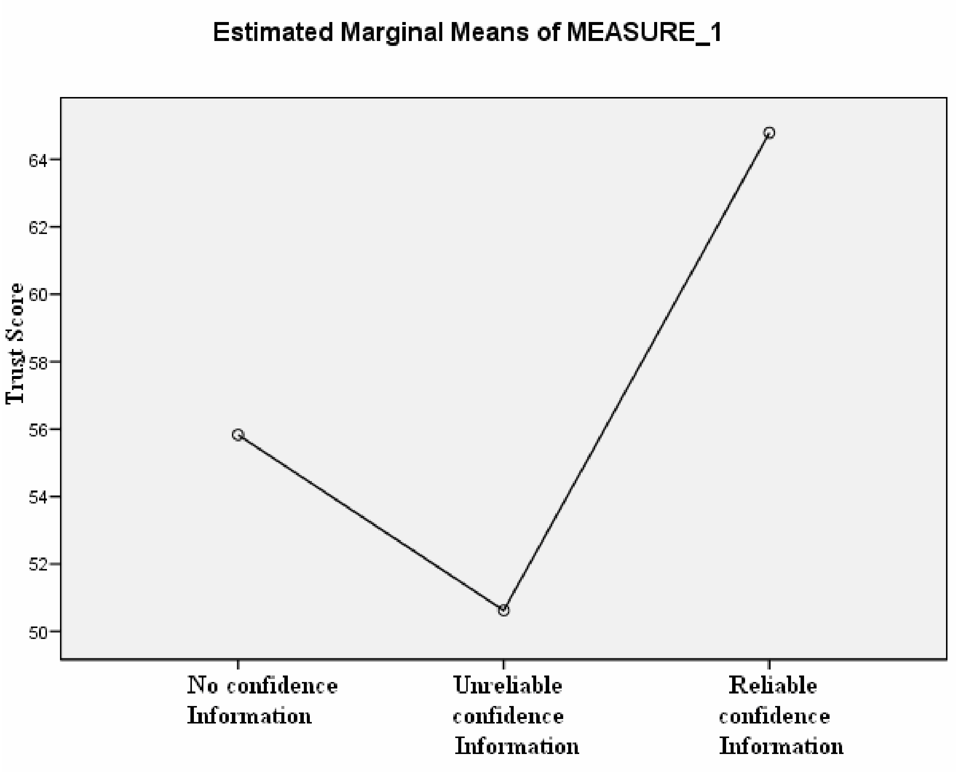


Figure 4-13: Marginal Mean of the three "confidence information" level on trust score of 75% accuracy group

Table 4.2 is based on a results obtained form the reliance questionnaire in all six automatic modes. The functionality types (Spot or track) which led to better reliance scores in each of the five factors is shown in the second column on the Table 4.2. The "confidence information" level which has higher scores in each of the five factors is shown in the third column of the table. Those marked with an asterisk in the "confidence information" level column were found to be significantly different comparing to the other two "confidence information" levels,

Reliance factor	Functionality level	Confidence information Level
Effectiveness	Track	Reliable confidence information
Trust	Track	Reliable confidence information*
Risk	Spot	Unreliable confidence information*
Confidence	Track	No confidence information
State of Learning	Track	Reliable confidence information

Table 4-2 : highest scores for the five reliance factors in 75% accuracy

4.2.3 Comparison of the 66% and 75% Accuracy Levels

In this section the total scores of RTLX, secondary task performance and reliance factors' scores in the two accuracy groups have been compared. This was done using a between subject independent sampled t-test. The hypotheses tested in this part of the experiment are:

H1: There is a significant difference between the RTLX score of 66% and RTLX score of 75% automated modes.

H0: There is no significant difference between the RTLX score of 66% and RTLX score of 75% automated modes.

The RTLX scores of the two groups have been tested and the only mode which showed significant difference between the two accuracy levels was "no confidence spot" ($P<0.05$, 2 tailed). The RTLX score for the "no confidence spot" at the 66 % accuracy level was significantly higher (mean=59.23) than RTLX score in the 75% level (mean=46.11).

H1: There is a significant difference between the secondary task performance score of 66% and secondary task performance score of 75% automated modes.

H0: There is no significant difference between the secondary task performance score of 66% and secondary task performance score of 75% automated modes.

Secondary task performance of the two accuracy groups has been tested; no significant difference was observed among any of the automated modes between the two accuracy levels.

H1: There is a significant difference between the reliance factors' score of 66% and reliance factors' score of 75% automated modes.

H0: There is no significant difference between the reliance factors' score of 66% and reliance factors' score of 75% automated modes.

Testing the five reliance factors against the hypotheses revealed that only "effectiveness" scores in the "No confidence Track" mode of the 66% accuracy level is significantly different from "effectiveness" scores in the "No confidence Track" mode of 75% accuracy. The 66% accuracy level has a better effectiveness score in the "No confidence Track" mode (mean=70.00) than the effectiveness score of "No confidence Track" mode in the 75% accuracy (mean=60.20).

4.3 Comparison of Automated and Baseline Results

RTLX and secondary task performance scores for the experimental automatic and baseline systems have been compared by a within subject paired sample t-test. The pairs used for investigating the hypothesis listed below were: (No confidence Spot, baseline), (No confidence Track, baseline), (Unreliable confidence Spot, baseline), (Unreliable confidence Track, baseline), (Reliable confidence Track, baseline) and (Reliable confidence Track, baseline).

H1: There is a significant difference between RTLX score and secondary task performance score of the automatic systems and RTLX and secondary task performance score of baseline.

H0: There is no significant difference between RTLX score and secondary task performance score of the automatic systems and RTLX score and secondary task performance score of baseline.

4.3.1 66% Accuracy Level

The results show a significant difference between secondary task performance score of baseline and "Unreliable confidence Spot" ($P<0.05$); baseline and "Reliable confidence Spot" ($P<0.05$); baseline and "Unreliable confidence Track" ($P<0.05$); and baseline and "Reliable confidence Track" ($P<0.001$). No other significant differences were observed ($P>0.05$).

4.3.2 75% Accuracy Level

The results show that there is a significant difference between secondary task performance score of baseline and “Unreliable confidence Spot” ($P<0.01$); baseline and “Reliable confidence Spot” ($P<0.05$); baseline and “No confidence Track” ($P<0.01$); and baseline and “Reliable confidence Track” ($P<0.001$). No other significant differences were observed ($P>0.05$).

4.4 Audio Description Results

The methods adopted by participants to identify the target are different. Based on the audio description they had information regarding the height, weight, clothing, age, gender and ethnicity of the target. All of the participants used information about clothing as their main source of information, two out of the 24 students used height and weight, one of the participants noted that “weight and height was useless because I was not familiar with standard metrics and could not convert them”. Two of the participants used clothing for the first identification and then used gait which is the pattern of movement developed by body joints.

5 Discussion

The findings of this project suggest that providing information regarding the confidence of alarms for even an imperfect automatic surveillance system can improve the quality of the surveillance. Automatic surveillance systems are far from being satisfactory; however taking human operators into account during the design of even an imperfect surveillance system significantly increases the performance.

There are many studies in the literature which emphasise the role of the workload assessments for designing better human automation interaction. Surprisingly unlike auto pilots and automatic glass cockpits or automatic cruise controls, research on workload assessment for automatic surveillance system seems to be scarce. Similar studies in assessing workload of other automatic aids such as adaptive cruise controls were used for comparison of results obtained in this experiment.

This chapter aims to discuss the findings of this project and their implications. Section 5.1 discusses the results obtained from two baseline scenarios (both manual cases) in either of the accuracy groups. The extent to which the experimental automatic surveillance system differs from the manual surveillance is discussed in section 5.2. Section 5.3 notes which functionality type is preferred, section 5.4 discuss various levels of "confidence information". Finally accuracy of the automated surveillance systems is discussed in section 5.5.

5.1 Baselines' differences

Although the difference between the baseline scores of the two accuracy levels was not significant, the 75% accuracy level baseline video clip seems to reduce the workload demands and enhance the performance. However, the only difference between the 66% and

75% accuracy levels is that in the latter the number of appearances of the target increases. Initially the assumption was that since the target appeared once more than the 66% level baseline and because both video clips were displayed to the participants for similar lengths of time (i.e., two minutes), the operator's workload would be more the 75% level. The results of the experiment proved thus assumption to be wrong. This is due to the increase in the level of arousal. As Megaw (2005) has rightly pointed out, the level of arousal influences operators' workload and it can change the direct relationship of task demands and workload. However this does not mean that surveillance of crowded places will lead to better performance, since the two baselines are identical and the only difference is just an extra appearance of the target in the 75% accuracy level.

5.2 Automatic vs. Manual Experimental Surveillance Systems

Comparing the automatic system with the manual system, it became apparent that the [which system] did not significantly decrease subjective workload score (RTLX) for either of the accuracy groups. Nilsson, (1995) and Ward et al. (1995) obtained similar results from an automatic adaptive cruise controller. One of the possible reasons of the unchanged RTLX scores is due to the duration of each trial (two minutes) and the fact that reactive monitoring is not as tedious as proactive monitoring. If the duration of the trial was longer than two minutes then it might have affected the RTLX scores too.

However in some modes of the experimental automatic system, the workload score for the secondary task performance was increased. In other words, the spare attentional capacity for the automated mode is superior to the spare attentional capacity of the baseline. This finding is similar to results of an experiment performed by Stanton et al. (1997) where automatic cruise controls significantly improved the secondary task performances. The duration of each trial in the study reported by Stanton et al. (1997) was two minutes. Young and Stanton (2002) revised the Stanton et al's (1997) experiment by just changing the duration of trials to 10 minutes and found that the secondary task performance will not significantly differ in the automated modes.

When the system is 66% accurate, in both of the functionality levels (spot and track) it should give the some level of "confidence information" (unreliable or reliable) to increase the spare attentional capacity. This trend is slightly different in the 75% accuracy level; it seems that when the functionality of automatic system is set to “tracking”, regardless of the level of “confidence information”, spare attentional capacity is improved. On the other hand, when the functionality is “spotting”, if the system does not provide “confidence information” (No confidence Spot), the amount of spare attentional capacity will not increase, whereas in the other two levels of “Unreliable confidence Spot” and “Reliable confidence Spot” the spare attentional capacity is increased.

These results show that although automation does not change the subjective rating of workload compared to the manual monitoring, but it can improve the performance by providing operators with more spare resources.

5.3 What task should be prioritised?

Two automated tasks were investigated in this project: Spotting and Tracking. Although automated tracking systems gave better RTLX and secondary task score, they did not improve the performance significantly. As noted in Section 2.2.3, current tracking automated systems are more accurate. Perhaps one of the important results obtained from this experiment is that regardless of the accuracy of the automated tracking or spotting, operators find automated tracking more useful and less demanding. The reason is that when an automatic system spots the wrong target, the task demands will increase. Whereas in the automatic tracking mode, if there is any mistakes operator can simply ignore it because they have already spotted the true positive target.

Furthermore, in terms of the reliance factors as it was shown in tables 5.1 and 5.2, the automated tracking system was more effective, more trustworthy, and easier to learn, increased the confidence of operators and was less misleading and risky. Riley (1996) stated that by considering the reliance factors, the use of automation will be encouraged and this will offer the operators a good understanding of various facets of their decision making process.

The means of assessing reliance factors in this study was limited to a short questionnaire. This might explain the non significant effects; however since participants had to fill in two questionnaires after each trial, having a long, detailed questionnaire would cause frustration and would affect the whole experiment. These results suggest that further research on the reliance factors is required.

5.4 How helpful is an informative system?

One of the most important findings of this project might be the confirming the fact that providing operators with information about systems' confidence in the alarms, significantly decreases workload and increases spare attentional capacity. This was the case in both of the accuracy levels. Moreover in the context of automatic surveillance systems, there are many challenges to develop an ideal and fully functional automatic surveillance system. The results of this project suggest that with human-centred approach towards automation the effectiveness of the current technology can be increased. Providing feedback about the systems' reliability is a way of making status of the system more visible to users and hence more usable.

"Confidence information" also affects the reliance factors. Although in the 66% accuracy level none of the reliance factors were significantly affected, when the system gives no information about the alarms generated (No confidence information), the effectiveness score is higher. The reason might be that since there is no colour differentiation in this mode it is easier to learn..

Providing the operator with unreliable information about system's level of confidence can increase the risk, since it can be misleading. Finally when the system gives reliable information about the alarms generated, operators' subjective level of trust is higher than the other two modes.

5.5 How much accuracy is good enough?

The subjective workload score (RTLX) of automated spotting system, when it does not give any information (No confidence Spot), is significantly lower in the 75% accuracy group compared to the 66% accuracy level. In terms of secondary performance scores, although the amount of spare attentional capacity has increased in the 75% accuracy level compared to the 66% accuracy level, this improvement is not significant. This suggests that although the subjective workload score decreases due to the increase in accuracy, the performance of secondary task will not significantly deteriorate. Furthermore, for the reliance factors of each of the automated modes in both of the accuracy levels, the effectiveness of "No confidence Track" in 66% accuracy level was significantly better than the effectiveness of "No confidence Track" in 75% accuracy level.

Note must be taken that the accuracy of this experimental automatic system is only determined by the rate of true positives against the total detections. The accuracy of an automatic surveillance system is affected by many factors, as it has been described by PETS and FERET evaluation methods (Dee and Valestine, 2007). Simplification in this experiment might have caused some type II errors in cases that were reported non significant. But the only significant case was found strongly suggests that although increase in the accuracy has the potential to increase performance and decrease workload but it will not necessarily increase the subjective effectiveness of the automatic system, which might decrease the reliance on automation.

6 Recommendations and future work

This project generates a series of recommendations for the design of future automatic surveillance systems. Perfect automatic surveillance systems which are fully functional and accurate in real life situation seem to be out of rich for the time. However, current systems can simply improve by giving information regarding the reasons of automatic choices, similar to the (Dzindolet et al., 2003), to know why aid made errors or in this project to know what are the aid’s decisions are based on, will increase their trust in the aid.

Although the “unreliable confidence information” which is an imperfect feedback can cause confusion, it still increases operators’ performance. Applying this notion in the design of future automatic surveillance systems is feasible and it does not require any technological advancements.

Furthermore, considering the functionalities, automatic tracking systems seem to be more effective. Hence a system in which the operator detects the target manually and lets the system to continue tracking and reporting targets’ destination seems to be helpful. Note must be taken that this recommendation is for moderately busy locations where operators can detect better than the system. It is necessary to investigate the validity of thus recommendation for highly crowded locations.

The increase in the accuracy of the system will not necessarily lead to a better performance. Although only a very simple implication of accuracy has been applied in this project, it might suggest that a 9% increase in the accuracy, the transition from 66% to 75%, will not significantly affect the performance, the question of the extent to which higher levels of accuracy (higher than 75%) affect the performance of automatic surveillance systems needs to be addressed. Limitations towards the increase of accuracy in automatic surveillance systems as mentioned in Chapter two re-emphasise the importance of focusing on improving performance with a way which is independent from vision algorithms’ efficiency.

Further research in various dimensions of human automation interaction can elaborate more rigorous guidelines for the design of future automatic surveillance systems. This study did not consider the effect of skill and all of the participants are novice operators. Future research can investigate how expert operators interact with automatic surveillance systems. As mentioned earlier, the automatic system tested in this study was a very simplified mock up with a straightforward definition of accuracy; it will be interesting to design a more functional robust prototype of automatic surveillance system with more distant levels of accuracy.

Automation has long been a tool to maintain the situation awareness of operators in dynamic environments. There is a relationship between situation awareness and mental workload as described in (Warm., et al., 1997). An ideal system is one which increases operators' situation awareness and decreases his/her mental workload and these should not be considered independent from each other. In the current project these two concepts are connected to each other. If the operator's mental workload somehow decreases through automating some aspects of the task, he/she can spent more time and attention on other tasks and consequently his/her situation awareness will increase.

References

[1].Ainsworth, T.: Buyer Beware. In: Security Oz, **19**, pp.18-26 (2002)

[2].Armitage, R.: To CCTV or not to CCTV? A review of current research into the effectiveness of CCTV system in reducing crime. NACRO, June 2002.

[3].Black, J., Velastin, S., Boghossian, B.: A Real Time Surveillance System for Metropolitan. In: Proceedings of. International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 189–194, (2005)

[4].Bliss, J.: An investigation of alarm related accidents and incidents in aviation. In: International Journal of Aviation Psychology, **13**, pp.249-268, (2003)

[5].Boghossian, B., Black, J.: The challenges of robust 24/7 video surveillance systems. In: IEE International Symposium on Imaging for Crime Detection and Prevention, ICDP, pp.33-38, (2005)

[6].Byers, J.C., Bittner, A.C., Hill, S. G.: Traditional and raw task load index (TLX) correlations and paired comparison necessary? In: Advances in Industrial Ergonomics and Safety I, edited by Mital, A. Taylor & Francis, London (1989)

[7].Cameron, H.: CCTV and (In)dividualtion. In: Norris, McCahill and Wood (eds.) Surveillance and Society, CCTV Special, pp.136-144. (2004)

[8].Davies, A., Yin, H., Velastin, S.: Crowd monitoring using image processing. In: Electronics & Communication Engineering Journal, (1995)

[9].Dee, H.M., Velastin, S.: How Close are we to solving the problem of automated visual surveillance? In: Machine Vision and Applications, (2007)

[10]. Dick, A.R., Brooks, M.J.: A Stochastic Approach to Tracking objects Across multiple cameras. In: Advances in Artificial Intelligence, pp. 160-170, (2004)

[11]. Dixon, S., Wickens, C.S.: Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload. In: Human Factors, **48**, pp. 474-486, (2006)

[12]. Donald, C.H.M.: Assessing the Human Vigilance Capacity of Control Room Operators. In: people in Control: An International Conference on Human Interfaces in Control Rooms, IEE (1999)

[13]. Donald, C.: Vigilance .In: People in Control: human factors in control room design, edited by Noyes, J., Bransby, M., IEE Control Engineering Series 60, (2001)

[14]. Dzindolet, M.T., Peterson, S.A. Pomranky, R.A., Pierce, L.G., Beck, H.P.: The role of trust in automation reliance. In: International Journal of Human-Computer Studies. **58**, pp.697-718, (2003)

[15]. Endsley, M.R.: Toward a theory of Situation Awareness in Dynamic Systems. In: Human Factors: The Journal of the Human Factors and Ergonomics Society, Vol 37, pp.32-64, (1995)

[16]. Ferryman, J., Adams, A., Velastin, S., Ellis, T., Remagnino, P., Tyler, N.: Case for Support, Part1: Previous Research and Track Record .In: REASON: Robust Methods for Monitoring ands Understanding People in Public Places, (2008)

[17]. Hampapur, A., Brown, L., Feris, R., Senior, A., Shu, Ch., Tian, Y., Zhai, Y., Lu, M.: Searching Surveillance Video. In: Advanced Video and Signal Based Surveillance (AVSS) IEEE, pp. 75-80, (2007)

[18]. Harris, S. Stedmon, A. Sharples, S. Wilson, J.: Keeping Track of a Suspect on CCTV: How Hard Can It Be?! In: P.Bust (ed.)Contemporary Ergonomics, Taylor & Francis, London, (2008)

[19]. Hart, S.G., Staveland, L.E.:Development of NASA-TLX(Task Load Index): Results of Empirical and Theoretical Research. In: P.A. Hancock, N.Meshkati (eds.), Human Mental Workload, pp.239-250 , Amesterdam, North Holland Press, (1988)

[20]. Hayman, J.: Computer Vision Based People Tracking for Motivating Behaviour in Public Spaces. In: Massachusetts Institute of Technology, (2003)

[21]. Hogan, J.: Your every move will be analysed. In: New Scientists Magazine. Issue 2403, (2003)

[22]. Jamieson, G.A.: Developing Human-Machine Interfaces to support Appropriate Trust and Reliance on Automated Combat Identification Systems, Progress Report for Milestones 1,2 and 3, (2007)

[23]. Keval, H.: CCTV Control Room Collaboration and Communication: Does it Work? In: Proceedings of Human Centred Technology Workshop, 2006.

[24]. Kantowitz, G.H., Campbell, J.L.: Pilot Workload and Flight deck Automation. In: R.Parasuraman, M.Mouloua (eds.) Automation and Human performance, Theory and Applications, pp. 117-137, (1996)

[25]. Lee, Y.H., Liu, B.S.: In flight Workload Assessment: Comparison of Subjective and Physiological Measurements. In: Aviation, Space and Environmental Medicine, **74**, pp. 1078-1084, (2003)

[26]. Lo, B.P., Sun, J., Velastin, S.: Fusing Visual and Audio Information in a Distributed Intelligent Surveillance Systems for Public Transport Systems. In: ACTA AUTOMATICA SINICA, **29**, (2003)

[27]. Luff, P., Heath, C.: Naturalistic analysis of control room activities, In: J.Noyes, M.Bransby (eds.) People in Control, Human factors in control room design, IEE Control Engineering Series 60, (2001).

[28]. McKenna, S., Gong, S.: Non-intrusive Person Authentication for Access Control by Visual Tracking and Face Recognition. In: Proceeding of the First International Conference on Audio- and Video-based Biometric Person Authentication, pp. 177-183, (1997)

[29]. Megaw, T.: The definition and measurement of mental workload. In: Evaluation of Human Work, Edited by Wilson, J.R., Corelt, N. Third edition, Taylor & Francis, (2005)

[30]. Moeslund.T and Granum.E, A survey of Computer Vision-Based Human Motion Capture, Computer Vision and Image Understanding 81, 231-268(2001)

[31]. Mosier, K.L., Skitka, L.J.: Human Decision Makers and Automated Decision Aids: Made for Each Other? In: R.Parasuraman, M.Mouloua (eds.) Automation and Human performance, Theory and Applications, pp. 201-218, (1996)

[32]. Muller-Schneiders, S., Jegar, T., Loos, H.S., Niem, W.: Performance Evaluation of a Real Time Video Surveillance System. In: Proceedings of 2nd Joint IEEE International Workshop on VS-PETS. Beijing, (2005)

[33]. Neil, D., Thomas, N., Baker, B.: Threat Image Projection in CCTV. In: Security Technology, 41st Annual IEEE International Carnahan Conference, pp.272-280, (2007)

[34]. Norris, C., McCahill, M.: CCTV: Beyond Penal Modernism? In: Brit.J.Criminol. pp. 97-118. (2006)

[35]. Park, K.S., Yun, T.D., Kim, C.H., Lee, S.S.:Comparison of the task performance with the multi-display computer. In: Proceeding of Fifth International Conference on Computational Science and Applications. IEEE, (2007)

[36]. Riley, V.: Operator Reliance on Automation: Theory and Data. In: R.Parasuraman, M.Mouloua (eds.), Automation and Human performance, Theory and Applications, pp. 19-35, (1996)

[37]. Sage, K., Young, S.: Security Applications of Computer Vision. In: Aerospace and Electronic Systems Magazine, IEE, pp.19-29, (1999)

[38]. Sanders, M.S., McCormick, E.J.: Human Factors in Engineering And Design, Sixth Edition. McGraw-Hill Book Company, (1987)

[39]. Sanderson, C., Bigdeli, A., Shan, T., Chen, Sh., Berglund, E., Lovell, B.: Intelligent CCTV for Mass Transport Security: Challenges and Opportunities for Video and Face Processing. In: Electronic Letters on Computer Vision and Image Analysis, **6**, pp.:30-41, (2007)

[40]. Shan, T., Chen, Sh., Sanderson, C., Lovell, B.: Towards robust face recognition for intelligent CCTV based surveillance suing one gallery image. In: Advanced Video and Signal Based Surveillance (AVSS), IEEE, pp.470-475, (2007)

[41]. Shneiderman, B.: Eight Golden Rules of Interface Design (third ed.), In Designing the User Interface, Addison-Wesley, London (1997)

[42]. Stanton, N.A., Salmon, P.M., Walker, G.H., Baber, C., Jenkins, D.P.: Human Factors Methods, A Practical Guide for Engineering and Design. ASHGATE, (2005)

[43]. Stanton, N.A., Young, M., McCaulder, B.: Drive-By-Wire: The Case of driver Workload and Reclaiming Control with Adaptive Cruise Control. In: Safety Science, **27**, pp. 149-159, (1997)

[44]. Sauer, J.: Prospective memory: a secondary task with promise. In: Applied Ergonomics **31**, pp. 131-137, (2000)

[45]. Valera, M., Velastin, S.A.: Intelligent distributed surveillance systems: a review. In: Vision, Image and Signal Processing, IEE Proceedings, pp.192-204, (2005)

[46]. Velastin, S.A, Boghossian, B.A., Vicencio-Silva, M.A.,: A motion-based image processing system for detecting potentially dangerous situations in underground railway stations. In: Transportation Research Part C, pp. 96-113. ELSEVIER, (2006)

[47]. Ward, N.J., Fairclough, S., Humphreys, M.: The effect of task automation in the automotive context: a field study of an autonomous intelligent cruise control system. In: International Conference on Experimental Analysis and Measurement of Situation Awareness, pp. 369-374, (1995)

[48]. Warm, J. Dember, W., Hancock, P.: Vigilance and workload in automated systems. In: R.Parasuraman, M.Mouloua (eds.), Automation and Human performance, Theory and Applications, pp. 183-199, (1996)

[49]. Wallace, E., Diffley, C.: CCTV Making It Work: Control Room Ergonomics. In: Police Scientific Development Branch, Home Office, St Albans (1998)

[50]. Wickens, C.D., Lee, J.D., Liu, Y., Gordon Becker, S.E.: An introduction to Human Factors Engineering (Second edition), Pearson Education, Inc., (2004)

[51]. Young, M.S., Stanton, N.A.: Taking the load off: investigations of how adaptive cruise control affects mental workload. In: Ergonomics, **47**, pp.1014-1035, (2004)

Electronic References

[1]. http://dilnxsrv.king.ac.uk/cromatica/FinalReport.pdf , Accessed on [29/06/08]

[2]. www.detec.no, Accessed on [29/06/08]

[3]. www.petsmetrics.net, Accessed on [29/06/08]

[4]. http://www.frvt.org/ , Accessed on[29/06/08]

[5]. http://www.rail-reg.gov.uk/server/show/nav.1529 Accesses on [5/07/08]

Appendices

Appendix I: Counter Balancing Sheet

Key:

NT→ No confidence Track

NS→ No confidence Spot

US→ Unreliable confidence Spot

UT→Unreliable confidence Track

RS→ Reliable confidence Spot

RT→ Reliable confidence Track

NT	US	RS	RT	UT	NS
US	RT	NT	NS	RS	UT
RT	NS	US	UT	NT	RS
US	UT	RT	RS	US	UT
UT	RS	NS	UT	RT	US
RS	NT	UT	US	NS	RT

Appendix II: Experiment Questionnaires and forms

Consent form:

I, Nastaran Dadashi would like to thank you for your participation in this experiment; it is a final part of my Masters course, MSc Interactive Systems Design in School of Mechanical, Materials and Manufacturing Engineering and School of Computer Science, under the supervision of Dr.Alex Stedmon and Dr.Tony Pridmore.

In this experiment which will take about an hour you will be asked to perform eight sets of observational and monitoring tasks each contains video clips of two minutes length. After briefing about the experiment you will be asked to fill in two questionnaires, the result of the questionnaire will then be analysed by statistical softwares.

Participation in this experiment is strictly voluntary and you may withdraw at any point. Your participation in this experiment is anonymous and the records of your participation will be kept strictly confidential.

If you felt any inconvenience at any stage of the experiment please make sure that you inform me.
For any questions regarding the experiment please do not hesitate to ask me.

Participation Consent

I confirm that I have read and understood the above information and description of the study and agree to take part

Signed……………………………………………………..Date………………………..

Information sheet:

Title of Project:

Reducing workload in CCTV monitoring

Researcher name and contact information

Nastaran Dadashi, psxnd@nottingham.ac.uk

Supervisor's name and contact information

Dr. Alex Stedmon (Alex.Stedmon@nottingham.ac.uk)

Dr.Tony Pridmore (tpp@Cs.Nott.AC.UK)

Aim of the project

The aim of this project is to understand whether automated surveillance system can improve operator's performance, mainly the two aspects being analysed in this project are functionality of the automatic system and the level of confidence operators have in such automatic systems.

Participation in the study

Participation in this study is voluntary and participants are able to withdraw at any time, name or any specific demographic information is not required for this project. The researcher will be available during the research phase of the study to answer any questions participants may have.

Use of study findings

No individual data will available to anyone but the researcher. All individual data collected will be kept confidential, the data collected will then be analysed a statistical software. The identity of participants will keep anonymous in the report of the results.

Pre-experimental test:

Do you recognise any of the persons in the pictures? Please tick as many as appropriate.

Results sheet:

	Spot right	Track right	Missed Mistake	Ignored right	Back pack	Comments
S						
T						
NS						
NT						
US						
UT						
RS						
RT						

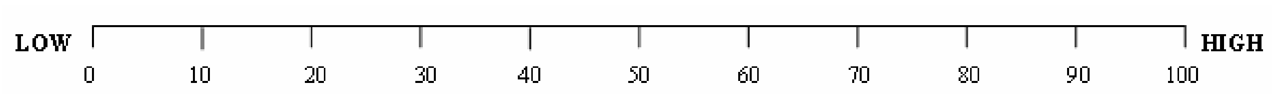
NASA-TLX pro forma:

Mental Workload Questionnaire (NASA-Task Load Index) Task code…………….

This questionnaire is designed to assess the level of difficulty you associate with the tasks you performed in the experiment. Please rate the level of difficulty by placing marking the appropriate point anywhere on each line to indicate your workload estimate.

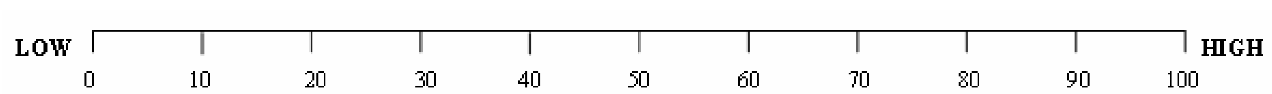
1) Mental Demand

How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, forgiving or exacting?



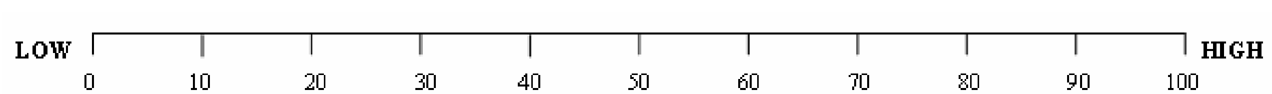
2) Physical Demand

How much physical activity required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?



3) Temporal Demand

How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?



4) Performance

How successful do you think you were in accomplishing the goals of the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

LOW 0 10 20 30 40 50 60 70 80 90 100 HIGH

5) Effort

How hard did you have to work (mentally or physically) to accomplish your level of performance?

LOW 0 10 20 30 40 50 60 70 80 90 100 HIGH

6) Frustration Level

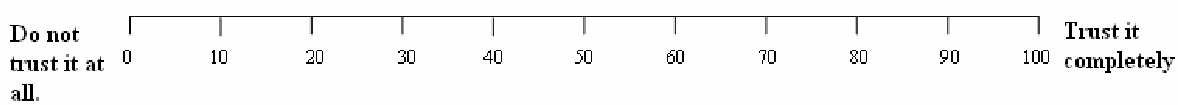
How insecure, discouraged, irritated, stressed and annoyed(high levels of frustration), versus secure, gratified, content, relaxed and complacent (low levels of frustration) did you feel during the task?

LOW 0 10 20 30 40 50 60 70 80 90 100 HIGH

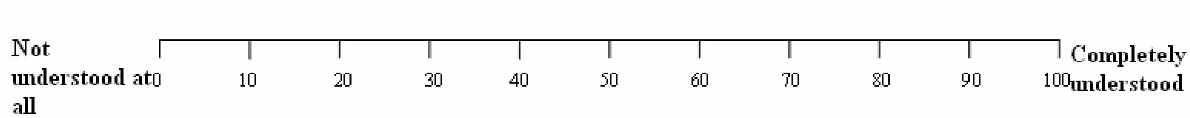
Reliance Questionnaire:

Please answer to the following questions by marking anywhere on the scale below.

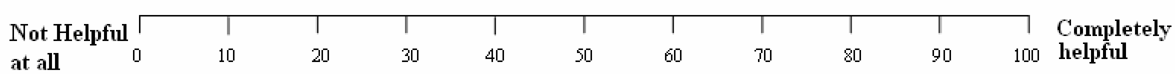
1. How much trust you have in the system?



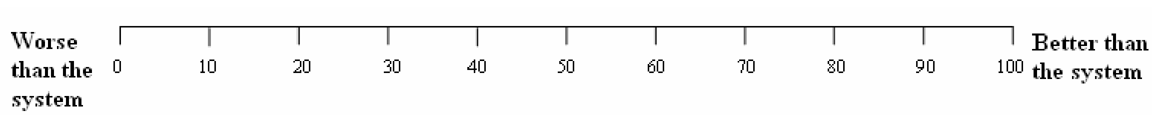
2. How did you understand the information presented to you.



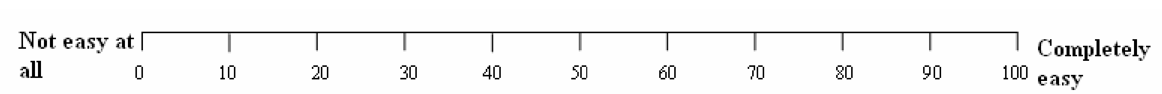
3. How helpful did you think the automatic system was?



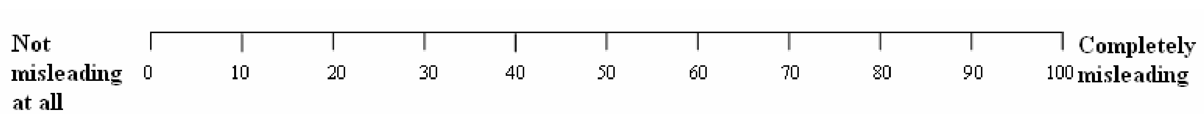
4. How can you spot/track the suspect comparing to the automatic system?



5. How easy was the system to use?



6. How misleading the automatic system was?



Participation Record

I ……………………………………confirm that I have completed viewing the eight video clips as required in the experiment .I understand that upon submission the expense form I shall receive ₤ 8 for completion of the experiment.

Signed ………………………………………………………………..Date………………….