UNCLASSIFIED

Human and Algorithm Facial Recognition Performance: Face in a Crowd

Master of Science
Stacy, Emily
[July 2017]

Australian Government
Department of Defence
Defence Science and Technology Organisation

UNCLASSIFIED
In partnership with the Defence Science and Technology Organisation

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student

Date

Acknowledgements

Firstly, I would like to express my sincere gratitude to my external supervisor Dr. Brett McLindin for the continuous support of my Masters research, for his constant patience, tenacious motivation and immense and unmatched knowledge. His guidance has helped me through the entirety of research and writing this thesis and I could not have imagined having a better advisor and mentor.
I would also like to thank my UTS supervisors Dr. Meiya Sutisno and Professor Allan Jones, for their support and guidance of my research.

Alongside my three supervisors, I would like to thank the Defence Science and Technology Organisation for allowing me the privilege of joining their team as a student, and giving me access to their laboratory and research facilities. Without their support this research would not have been possible.
Also thanks goes to the University of Technology Sydney for granting me the opportunity to undertake research in a partnership with an external agency.

I thank my fellow students and colleagues at Defence Science and Technology Organisation for their support and willingness to help, especially during the operational trial.

Last but not least, I would like to thank my family, and mainly my Mum, Robyn. She has continuously supported and encouraged me and I feel there are not enough words to use to express my gratitude. This thesis would not be complete if it was not for her.

Abstract

Developing a method of identifying persons of interest (POIs) in uncontrolled environments, accurately and rapidly, is paramount in the 21st century. One such technique to do this is by using automated facial recognition systems (FRS). To date, FRS have mainly been tested in laboratory conditions (controlled) however there is little publically available research to indicate the performance levels, and therefore the feasibility of using FRS in public, uncontrolled environments, known as face-in-a-crowd (FIAC). This research project was hence directed at determining the feasibility of FIAC technology in uncontrolled, operational environments with the aim of being able to identify POIs. This was done by processing imagery obtained from a range of environments and camera technologies through one of the latest FR algorithms to evaluate the current level of FIAC performance. The hypothesis was that FR performance with higher resolution imagery would produce better FR results and that FIAC will be feasible in an operational environment when certain variables are controlled, such as camera type (resolution), lighting and number of people in the field of view. Key findings from this research revealed that although facial recognition algorithms for FIAC applications have shown improvement over the past decade, the feasibility of its deployment into uncontrolled environments remains unclear. The results support previous literature regarding the quality of the imagery being processed largely affecting the FRS performance, as imagery produced from high resolution cameras produced better performance results than imagery produced from CCTV cameras. The results suggest the current FR technology can potentially be viable in a FIAC scenario, if the operational environment can be modified to become better suited for optimal image acquisition. However, in areas where the environmental constraints were less controlled, the performance levels are seen to decrease significantly. The essential conclusion is that the data be processed with new versions of the algorithms that can track subjects through the environment, which is expected to vastly increase the performance, as well as potentially run an additional trial in alternate locations to gain a greater understanding of the feasibility of FIAC generically.

List of Figures:

List of Tables:

1. Introduction

The ability to accurately identify other human beings in an increasingly populous and complex world is one of the fundamental challenges of our age. Whether for personal reasons such as opening a bank account or applying for a passport, for public reasons such as travelling across a border, or for law enforcement purposes such as missing, wanted and/or excluded persons, the requirement to identify oneself and others is an important aspect of today's society. This thesis outlines the feasibility of using automated facial recognition (FR) as a means of identifying unknown persons in an uncontrolled environment, also known as Face in a Crowd (FIAC).

The method in which to identify humans has evolved over time, with the introduction of computer technology increasing the speed and efficiency that it can be done. Of particular focus for the current research is identification in applications that are uncontrolled, such as border security, surveillance, counter terrorism, casinos, missing and/or wanted persons and fugitives. In these scenarios, the one common goal is to be able to identify a person of interest (POI), which most often requires the comparison of a large number of people in an uncontrolled environment to a watchlist of previously known individuals or POIs.

The method to identify POIs in these conditions has predominantly focused on the face, as this is often the only feature available for non-intrusive, contact-free identification purposes. The face also allows identification with no cooperation from the person being imaged, which is particularly useful for surveillance scenarios [1]. Furthermore, facial images are the only form of identification available for some of these people and scenarios as, the face has historically been the most commonly used form of identification for identification documents (ID) (such as driver's license, passport). Traditionally the identification of POIs has been conducted by human operators, whereby trained personnel would monitor multiple closed circuit television (CCTV) screens and search for the POI [2]. However, Hampapur et al. and Bigdeki et al. (2003) have shown that human monitoring can be unreliable [3, 4]. Some studies have shown that humans can only remember a limited number of unfamiliar faces for identification and they become ineffective after a short period of time [5]. More specifically, results show operators' attention span decreases significantly when performing mundane tasks, becoming ineffective after 20 minutes [2, 6]. Further studies have also compared human performance to automated algorithms, finding that since 2003 automated algorithms have outperformed humans in facial matching tasks [7, 8].

Automated means of distinguishing between individuals based on a persons' biological characteristics is known as their 'biometric'. A biometric is "any automatically measureable, robust and distinctive physical characteristic or personal trait that can be used to identify an individual or verify the claimed identity of an individual" [9]. This can be based on physiological characteristics such as the face, iris, finger, hand or voice and/or behavioural traits such as gait, signature and others [10]. Although there are a number of useful applications for biometric systems based on behavioural traits, there are studies that argue that these traits are too subjective for identification purposes [11]. As a result, the focus of this study will concentrate on physiological characteristics as a means to distinguish between people in order to determine its feasibility.

The range of physiological characteristics used in biometric applications differ in the level of uniqueness, collectability and performance, and their use is dependent on the specific application in which they are intended. Finger and iris biometrics offer high uniqueness and good performance (low error rates) [12, 13] when the POI is actively participating in the identification process. However, in less controlled scenarios such as a crowd, where the POI is not necessarily aware of, or actively participating in the identification process, neither finger nor iris are able to be imaged reliably and hence used for biometric identification [4]. In these cases, the face is often the only identifiable feature that is accessible for contact-free identification, and as such, the use of the face in these scenarios has been coined FIAC. It is for this reason that the face and hence FR were chosen as the most likely biometric to be used to identify POIs in uncontrolled environments and used as the basis of this research.

Although FR works well in controlled scenarios where the person is participating in the process, there is a need for FR in uncontrolled FIAC scenarios where the POI is present in a public, uncontrolled space. In evaluating FR technology in relation to FIAC applications it is important to look at where the technology currently stands. There is literature spanning many disciplines that have researched the reported problems associated with FR when used in uncontrolled environments [14-17]. Research groups have reported that the FR technology is dependent on image quality, finding that better quality footage results in more accurate matching [16]. This can be problematic when the only available imagery for analysis has been acquired from CCTV, as is the case with FIAC scenarios. This is because more often than not, CCTV cameras have been installed to record imagery of a large, open area for security purposes and not specifically to acquire faces. Research has found that imagery with faces that are non-frontal with off-centre orientation, varying angles, facial poses and expressions combined with uncontrolled environmental factors (lighting, illumination and image resolution) results in poor matching result [6, 18]. Although there are a number of evaluations of FR tested in laboratory conditions (controlled) [8, 19], there is little publically available literature to demonstrate the performance of FR in uncontrolled operational evaluations.

The performance of FR in operational environments was demonstrated in a real life event in the recent 2013 Boston marathon bombings. The media viewed automated FR as a failure after the systems failed to identify the two suspected perpetrators though both offenders' had photos previously recorded in official government databases [20].

In an attempt to evaluate current FR technology and the potential reasons why these perpetrators were not identified, a study was conducted by simulating the scenario presented in the investigation. Two state-of-the-art facial recognition systems (FRS) were used to evaluate the maturity and readiness of FR in this FIAC application [21].The experimental setup used images of the suspects (published and released by law enforcement and news agencies) combined with one million mugshot images. The results from unfiltered facial searches found the suspects returning a true match score at varying rankings with some as low as 200,000+ positions. With results like this, it could be suggested that automated FR would not have worked in this situation, as human operators operating the FR technology would not scan matches placed that far in the returned matching ranks.

Results suggested that even though 1:1 matching generates highly accurate performance rates (when comparing two good quality images), when pose, illumination and expression are varied (as is the case with FIAC) performance rates drop [21]. Ultimately, the FBI resorted to showing the public the suspect images to ask for identification information as their FRS could not determine a match [20]. The main challenges in the case of the Boston bombings, which are the challenges in most general FIAC environments, were the poor lighting and off-axis image of the acquired face. The resolution of the imagery was also very poor as the CCTV camera had low resolution and slow frame rate that only managed to acquire a frame every few seconds which resulted in blurred, off-angle images. As part of the study, they found that the FR systems would not be ready for a "lights out" deployment with suggestions that more research/development was needed, with focus on coping with variations in pose, illumination and expression. The research did suggest that with the addition of demographic filtering and multiple probe images of the one POI, state-of-the-art face matchers could potentially assist law enforcement with identification using the face. This however, is more of a static FR tool and not a FIAC application.

There are a number of commercial companies and government institutions that are attempting to understand and overcome the technical and environmental factors that are known to affect FR from performing in FIAC applications. The Defence Science and Technology Organisation (DSTO) trialled the feasibility of FIAC technology both in an operational and scenario environment via two studies conducted in 2001 and 2006 [22, 23]. It was reported that the computer performance in 2001 was too slow for FIAC technology to work in real-time, finding that by the time the FR software had detected a match, the POI had departed the scene. As a conclusion, the report stated the technology, although promising was not developed enough to be deployed into an operational environment as a "set and forget" technology, and needed further development and testing [22]. The scenario trial in 2006 produced more promising results showing the technology may soon be feasible as an operational tool in controlled conditions. Although, this trial had actively participating subjects (POIs) which is not a FIAC application [23], meaning that further research would be required to determine feasibility of the technology in operation.

Further to this, a recently released study by the U.S Homeland of Security on their crowd-scanning FRS referred to as Biometric Optical Surveillance system (or 'BOSS') reveals workings toward improving the accuracy of automated facial identification in uncontrolled environments, and more specifically, at a distance [24]. The described aim was to develop an automated system that could actively or passively acquire a facial image from distances of up to 100 meters and be able to compare the image against a biometric database in less than 30 seconds. The system works to overcome the FIAC technical and environmental hurdles by creating a system with two towers, each bearing a robotic camera structure that encompasses an infrared ability and distance sensor. The cameras produce a stereoscopic image, which is two images of the same object taken from slightly different angles to create a 3-D signature. These 3-D signatures have been said to have the potential for an improved FR matching ability [24]. Currently however, the BOSS system is not deemed operationally ready, as these parameters cannot be met in an uncontrolled FIAC environment with 80 to 90 percent identification accuracy [25].

Private companies that have developed a FR system for potential use in FIAC applications have actively been working to develop their algorithms to function in scenarios with less-than-ideal environments. The issue lies in testing, as private companies cannot get access to large datasets or operational testing opportunities to evaluate their FR programs in order to report 'actual' FIAC results. As such, the American Government in partnership with the National Institute of Standards and Technology (NIST) has funded independent testing programs in an attempt to help develop FR capabilities by creating a standardised forum for private companies, universities and other vendors to participate in. This will be discussed further in Section 2.3, however, it has provided an opportunity for independent vendors to test their algorithms and develop them further on a large dataset. The most recent test in 2013 performed tests with a range of imagery quality types, including poor quality datasets. Results have shown that when using poorly constrained images (which best reflects FIAC), identification miss rates are typically between two and five times higher than the same searches with high quality images. The results did show that FR algorithms have improved accuracy when matching with non-ideal imagery from similar tests run in 2010; however the most recent test only examined poor quality images of cooperating participants. This therefore means that there is no current performance data of FR algorithms capability when using poor quality imagery with non-cooperative subjects, as is representative of FIAC applications.

With an increase in computer processing power since 2006 and vendors' efforts developing better algorithms, it could be argued that FIAC technology should again be tested in an uncontrolled operational environment to determine whether such an application is feasible as an operational tool in the current state.

Hence, the overarching problem posed in this thesis is to determine the feasibility of FIAC technology in an uncontrolled operational environment with the aim of being able to identify POIs. However, in doing this, the research is carried out by categorising and testing some of the variables that could affect FIAC performance. As such, there are four sub-problems posed as part of this research which, when combined, will give a greater insight into the performance and input into the assessment of the feasibility for FIAC deployment in operational environments.

The sub-problems are:

1. in an uncontrolled environment using existing CCTV cameras, what is the level of FIAC performance, and is it acceptable for operational deployment? (i.e. determining operational baseline performance);
2. if not, how does the operational level of performance compare when using CCTV cameras in a more controlled laboratory environment (best possible performance using existing cameras – laboratory baseline);
3. in a controlled environment using high resolution cameras, what is the level of FIAC performance? (laboratory ceiling performance); and
4. what is the level of FIAC performance in an operational environment using high resolution cameras? (operational performance with updated cameras).

It is hypothesised that the variables tested in the sub-problems will affect FR performance with higher resolution cameras and imagery producing better FR results. It is further hypothesised that FIAC will be feasible in an operational environment when certain variables are controlled for, including camera type (resolution), lighting and number of people in the field of view.

For each of these sub-problems the variable (camera/imagery) is tested using experimental methodology which follows in Section 3.1.

2. Methods of Human Identification

In order to determine the most appropriate means of identifying POIs in uncontrolled environments, and understand where the technology is currently positioned, it is important to examine the history and evolution of identification prior to focusing on FIAC scenarios.

The purpose of distinguishing between individuals in modern systems of identification has been a necessary task and problem since the late Middle Ages [26]. The necessity to determine ones identity dates back further still, with more medieval forms of identification such as branding and tattooing. However, more related to the current research was the introduction of "identification papers" and authentic documents such as the carrying of passports, which can be dated back to the fifteenth century. By the mid-sixteenth century, the requirement to hold a passport became a mandatory obligation, being considered a regulated means to verify identity. This contributed to the rise of new systems of registration through the official recording of information in the early modern world in 1792, with European countries no longer recording under local jurisdictions, but moving to a state registration system subject to central administration [27]. However the system was found to be fallible. It was becoming known that people could produce counterfeit documents that appeared to be authentic, which was a problem faced by medieval authorities and is still a problem in the 21st century as "individuals are identified by documents whose contents and use they cannot themselves determine" [26 p.253]. This led to the introduction of photographs in passports which was done in an effort to increase individuality and authenticity of paper documents, as humans could verify the resemblance between the person presenting the document and the image. This will be expanded on further in this chapter, however first it is important to differentiate between the meanings and applications of "distinguishing between individuals" in the 21st century.

Distinguishing between individuals can be for *verification* or *identification* applications. *Verification* asks "Is this X?" and occurs when the questioned identity needs to be authenticated against a known identity (or a group of known identities) [28]. In this scenario, Person X is compared in *one-to-one* (1:1) verification commonly used in access control where the person presents themselves and is compared against an ID [12]. An example of this is when a traveller presents themselves and their passport to a customs official for international travel. The customs official compares the live person against the image in the passport to determine whether 'Person

X' is who they say they are. Results from this type of comparison are a match or non-match and hence access to the controlled area being permitted or denied [29].

Alternatively, *Identification* asks the question of "Who is X?" and occurs when the identity is unknown. In this scenario, Person X is compared in a *one-to-many* (1:N) search where the goal is to identify an unknown person. This task is common in surveillance applications where, for example, the goal is to identify unknown persons such as missing or wanted people.

Identification and verification can be conducted by one or both of the following methods:

1. Identification by Humans
2. Identification by Automated means

2.1 Identification by Humans

The most common feature that humans utilise to distinguish between other humans in both *verification* and *identification* scenarios is the face. This is the case as most humans are not trained in alternate methods of identification that require expertise such as comparing fingerprint or iris pattern images. Additionally, the face is often the only easily visible, accessible, and identifiable feature available, and as such, is used for both *verification* and *identification* applications.

Unfamiliar face *verification* (such as used by customs officials) is predominantly conducted by comparing a previously acquired image (control image, e.g. passport) to the live person in a 1:1 comparison. Control images are typically acquired in optimal conditions where the face is frontal, the person has a neutral expression, the lighting is even and the background is neutral [30]. Although humans are typically good at recognising familiar faces, the ability to conclusively verify unfamiliar faces from images is a different process and requires a different set of skills, a task at which humans are typically poor [30]. This matching task has been simulated and assessed by a number of studies in which observers (participants) had to decide if two photographs showed the same persons' face or different people [31, 32].

Further studies reported field experiments that looked at *verification* in life interactions [33, 34]. The study by Kemp et al. (1997) examined the performance level of humans operating in *verification* tasks to accurately accept or deny the person tendering an ID. Results found that humans' performance was poor with more than 50% of the fraudulent IDs being incorrectly accepted as legitimate and about 10% of valid IDs being falsely rejected. Further studies by Bruce et al. (1999, 2001) looked at identification from video and CCTV in a study that represented a best-case scenario model. Human operators were assessed in their ability to verify whether a male person shown in a high quality 'target' image was present in a collection of 10 images depicting similar looking men. The images were all of clean shaven men acquired in controlled, laboratory conditions on the same day so as to create optimal matching conditions. The human participants performed poorly in this matching task, even though there were no time pressures or memory loading with the correct person only being identified in 70% of the cases. Additionally, when the study was reduced to a simple two-image match that asked the human operators to decide whether two images represented the same person, the error rate was 20% [35].

The results from these experiments averaged around 20-50% errors in person verification, even though the studies represent the best possible performance that could be obtained by a human operator. Realistically, operators working in the *verification* field are presented with images affected by time, often taken years apart with the live person possibly having a naturally altered appearance by factors such as weight gain/loss and/or ageing [36]. Real life verification is also often affected by time conditions where the operator is under pressure to make a decision [37].

The inaccuracy with which unfamiliar faces are verified can have dramatic consequences depending on the environment in which the task is being performed. Development in this field is of importance for border security, police and judicial scenarios [37]. It is for this reason that human verification tasks are being researched to possibly be improved upon by automated means.

I*dentification* applications, on the other hand, can range from comparing multiple images of what could depict the same person, to being able to identify POI in public in a 1:N search. In cases where a person has had an image acquired while committing a crime (such as from CCTV), a facial comparison expert would be required to compare the imaged POI against mug shot images of potential suspects. When comparing two facial images, human experts use a combination of methods such as anthropometry and morphology to determine whether the two identities are of the same person [38]. Anthropometry is the study of the human body measurements, as first introduced by Alphonse Bertillon in the nineteenth century, in an attempt to identify repeat criminal offenders [39]. It was believed that no two bodies were identical and so attempted to distinguish between them by recording and comparing their body measurements [38]. When applied to the face, some experts measure the distances between pre-determined facial landmarks and compare the measurements between the two images in question. It was noted that occasionally the same individual could return a different set of measurements which was a notable flaw in the ability of the system to correctly re-identify an individual [30]. There is much controversy surrounding this technique [34, 36] and as such may not be suitable for efficient identification.

Alternatively, another *identification* task is where personnel are required to identify potential POIs in open spaces. This is a difficult task as they are required to remember a number of unknown faces, and attempt to recognise and identify them when mixed with the public [40]. Traditionally, identifying POI has been conducted by trained human operators by monitoring multiple closed circuit television (CCTV) screens and searching for the POI [2]. This task is proven to be cognitively demanding and requires vigilant monitoring for rare events that may never occur as well as require an excessive working memory [41]. Research has shown that human monitoring is unreliable [3] with there being fundamental limitations in the facial recognition abilities of humans. Although humans are highly skilled at distinguishing familiar faces [42], they perform poorly when attempting to identifying unfamiliar faces [5, 43-45].

It has been argued that since humans can only remember a limited number of unfamiliar faces for identification [5] and that their performance decreases and becomes ineffective after 20 minutes [4, 6], tools such as automated FR may be able to assist in some regard. After the 9/11 bombings and cases of international rioting, thousands of CCTV cameras were installed in public places around the world in an attempt to firstly deter criminal activity, and secondly acquire it when it does occur. One of the major installations was in England, where it has been estimated that there are around 500,000 CCTV cameras in the London area and 4,000,000 cameras in the UK [6]. This data implies that there is approximately one camera for every 14 people. With this in mind, it would not be feasible to have enough security personnel to monitor the all camera feeds [46]. The most common method of analysing video footage to obtain identification information is *reactive* and offline. This refers to the analysis of forensic video post an event or incident, as was the case with the 9/11 terrorist attacks, London underground bombings and most recently the Boston Marathon Bombings. In the case of the 2005 London bombings, Law enforcement agencies were able to gather 6,000 hours of CCTV footage to reconstruct the events leading up to the disaster and potentially locate the POIs. It is not publically known whether there was attempts to examine this imagery with automated FR systems, however, with the studies cited earlier that found humans become inefficient after 20 minutes of performing facial identification, this extent of imagery would take an exorbitant length of time to forensically examine. As studies and history has shown, humans would not likely to have been able to identify the 'terrorists' involved in the aforementioned events from a crowd, and per se it has been suggested that FR technology possibly could have [47]. As such, it has become clear that there is a need for surveillance to be *proactive,* to work in real-time and to potentially prevent future incidents from occurring [48]. This could become possible with the aid of facial biometric automation with recent studies showing that, when performance levels of humans and computer algorithms were compared, algorithms have outperformed humans in facial matching tasks since 2003 [7, 8]. The question remains however, will this performance be an improvement in the operational environments and will the performance be good enough to warrant its implementation.

Additionally, research has shown that factors such as image quality, changes in appearance, and disguises can have an effect on human performance [49, 50]. Other factors found to affect the performance of humans includes: pose and illumination changes [51], and the 'other-race-effects' [52-54]. The 'other-race-effect' has been studied to show that humans have a tendency to be able to recognise faces from their own race more successfully than from other races [55, 56]. These studies and findings are of importance, as they highlight that the current means of *identifying* unfamiliar humans in uncontrolled settings is ineffective. As a result, it has become clear that *verification* and *identification* processes may be improved by the introduction of facial biometric automation [6].

2.2 Identification by Automated Means

As mentioned in Section 2.1, research has shown that humans can perform poorly when attempting to identify an unfamiliar person, and as such, efficiency and identification performance may be able to be improved upon with the assistance of automated systems. It has been argued that that a person's biological traits offer an effective means by which a person could be distinguished from another [12]. When these biological traits are processed by automated systems, they are referred to as a person's 'biometric'. Biometric traits are less likely to be misplaced, forged, forgotten or stolen when compared to other non-automated means of identification such as passwords, keys, tokens and/or cards [11]. This makes biometric traits the perfect candidates for verification and identification purposes.

2.2.1 Biometrics

As mentioned previously, biometrics is a general term that can be described as any measurable and/or distinctive physical characteristic or trait that can be used as a means of verification or identification against a claimed identity [9]. Biometric identification can be divided into physiological characteristics such as the face, iris, hand and voice or behavioural traits such as gait and signature [10]. The use of biometrics has been named as one of the 'top ten emerging technologies that will change the world' by the *Massachusetts Institute of Technology* (*MIT) Technology Review* [28]. Furthermore, a report released by the International Biometric Group (IBG) in 2009 claimed that the biometric industry was projected to grow from $3.42 million in 2009 to over $9 million in 2014 (http://www.ibgweb.com/) and reach $13.89 billion by 2017 [57]. Although there are benefits of being able to utilise behavioural traits, some research debates these methods are subjective and vulnerable to inconsistencies [58]. Thus, the focus of the current research will concentrate on physiological characteristics.

Human physiological characteristics offer a means to individualise humans from one another [59]. The study of these characteristics alongside biometric technology has found that they are unique to each person, providing the possibility of identification by discriminating between differences [58]. Of most focus in the recent years are the study of fingerprints, iris patterns and the face, which have now been described as the most accurate and technologically mature means of human identification which will be discussed below [57].

2.2.1.1 Fingerprints

Fingerprinting, technically termed Dactyloscopy is the study of the impression print left behind when the friction ridge patterns of human fingers and/or palms make contact with a surface [38]. The friction ridges found on fingers are raised portions of the epidermis that form in unique patterns on every human while developing as a foetus in the womb. Distinguishing between humans can be achieved by discriminating between these differences. Fingerprinting can be used for live *verification* applications and/or reactive forensic *identification* purposes. The term forensic is used as a reference to the application of scientific principles for use in the court of law. The process for forensic identification is reactive, which refers to the analysis of evidence after a crime has been committed, or post an event.

Traditionally, fingerprinting was a very manual process carried out by scientifically trained personnel who were educated on how to 'lift' a fingerprint using ink for forensic *verification* or *identification*. In the forensic situations, the trained personnel would dust for fingerprints at a crime scene, lift them using an ink method, and take an image of the print for comparison against other prints previously stored on file. The manual process of comparing countless fingerprints in order to verify or identify a person was slow, tedious, error prone and time intensive [60]. The introduction of computer technology in the 1960s meant the comparison of two or more prints could be automated. Automation allowed fingerprints to be scanned live (in real-time) and compared against a large repository of stored prints to either allow or deny access. The introduction of automated biometrics has combined the skills and techniques of humans with the processing power and memory of technology [60].

Fingerprinting is currently the leading biometric technology, with approximately 28.4% of the biometric market focused on ink fingerprinting and 38.3% of the market focused on Automated Fingerprint Identification Systems (AFIS). It is a proven technology that has a high level of accuracy determined by peer testing and the reproducibility of results [61]. The recorded error rates are low, especially when using AFIS, as this system has a function that alerts the user when the acquired print was of insufficient quality for matching purposes [62]. Fingerprint systems can include the use of peripheral devices, hand held units, imbedded devices, wall mounted versions and/or large units designed for heavy duty environments [63].

The sensors and processes that are used to acquire a fingerprint to process a match score can be small and low cost, as well as resistant to background lighting and temperature which allow this system to be deployed for a number of scenarios. Currently the United States of America (USA) fingerprints every traveller that enters their country and keeps their biometric on file for possible later use as part of their US-VISIT scheme. Fingerprinting is also being used daily for access to PCs, networks, restricted sites as well as being used as a means to authorise transactions [59]. Automated fingerprinting devices aid human analysts as they have the computing power and capability of comparing thousands of prints against one another, which would take human operators a considerable length of time, in a much shorter and efficient period of time.

Fingerprints however can only typically be used for limited purposes where there is physical contact with an object, which includes the identification of deceased, identification of forensic prints found at a crime scene and for verification applications in access control scenarios. Although there is research and development focused on contactless fingerprinting, for the purpose of this research, which is to identify a person in a crowd, fingerprinting, whilst accurate, is not currently feasible. Fingerprints are unlikely to be able to be acquired at a distance or without the active participation from the user and hence cannot be used for covert or surveillance scenarios. It is for this reason that fingerprints will not be researched further.

2.2.1.2 Iris Patterns

Another common physiological characteristic used for biometric verification is the iris pattern, which although is not a new research field, has had relatively new technological developments. These recent developments have flourished since 2005, when the iris identification patent that Flom and Safir held, as well as an automated iris recognition algorithm patent that Daugman held since the 1980's expired [8]. As a result, advancing computation capabilities has offered new avenues for iris comparison that humans previously could not do, as the iris pattern could not be detected by the human eye [59].

It is believed that the iris structure is unique and can be used as an identification tool to distinguish between individuals [64]. Iris patterns are illuminated by regular or infrared light, an image acquired, and then converted and stored as a template in a biometric system [63]. For the system to function, a high resolution image of the iris must be acquired with a high degree of participation from the user. The error rates for iris recognition are very low, with the Iris Challenge Evaluation in 2006 reporting an error rate (False Accept Rate) of just 0.001% [8]. Although the technological advances associated with iris recognition are showing vast improvement and possibility, the technology currently only allows for *verification* and limited *identification* scenarios where the applicant is actively participating in the process. This is due to the fact that the iris is very small (approximately 10mm in diameter) and current matching technology requires more than 200 pixels across the iris to be of "good quality". This would be difficult to acquire while the POI is 'on the move' and not directly cooperating [65].

As mentioned above, the focus of this research is to identify a person in a crowd, with the possibility of identification without any participation from the POI. Reports claim that with further development and technology advancement, 'iris-at-a-distance' and 'iris-on-the-fly', two methods possibly capable of scanning iris patterns at a distance will be capable of identifying a person covertly [65]. Hence, this technology requires more research and development before being feasible. However, the main driver for not using iris technology further for this research is the availability of iris images of POIs. Missing persons, and persons wanted by police are not likely to have iris images acquired for potential matching against the crowd. Conversely, facial images are commonly accessible and more likely to be available for comparison in these scenarios.

2.2.1.3 The Face

The face has been an area of interest in the automated biometric field since the 1960s as it allows for non-intrusive, contact-free identification [66]. The face also provides the benefit of acquiring a biometric 'sample', which is a biometric measure (face) presented to the system (FR algorithm) and acquired as an image (also known as a 'probe'), with no cooperation from the person, which is particularly useful for surveillance FIAC purposes. Furthermore, the face has advantages over fingerprints and iris patterns as there are different source types that can be used for enrolment, such as mug shot images, surveillance footage, live images and even images from printed media. Enrolment occurs when a sample (user) is loaded into a FR system, processed and generated into a template for storage and future identification purposes. The template is the stored sample that is then used as a basis for comparison to other templates

Humans use the face as a mechanism to distinguish between people on a daily basis. Although humans perform well when recognising familiar, known faces, as mentioned in 2.1, research has found that they perform poorly when attempting to identify unknown faces [5]. In *identification* tasks, without the aid of automated tools, a human would have to 'remember' a large number of unfamiliar faces in order to match them (1:N). Technology has been found to extend the human capability by being able to 'remember' and recall a large number of identities in a shorter period of time [66]. However, there are currently no automated systems that perform with high enough accuracy to allow a "set-and-forget" mechanism [30], and as such, still require the analysis and confirmation of humans to complete the process. In addition, studies have shown that humans can outperform algorithms in scenarios where information could be gathered from the environment that aid in human identification, such as gaining intelligence from backgrounds and utilising familiarity from aspects such as gait [40]. As a result, the work of O'Toole et al.

(2008) showed that by combining both the capabilities of humans and computerised systems together, it can lead to a near-perfect recognition system [67]. As such, the focus of facial identification has been on developing computer-assisted FR that can run in real-time to aid the human operator in identification and surveillance. It is for these reasons that the face was chosen as the focus of this study.

2.3 Facial Recognition

As photography and surveillance became more prevalent throughout the 20th century, so too did the study of methods that enable identification of the face from imagery [68]. Researchers have been developing FR systems since the 1960's [66, 69] with the focus and direction of the research shifting with the demands of society and law enforcement. What was originally focused on specific tasks such as basic facial detection and matching from still images, automated FR has evolved to its current state of a more 'end-to-end' product through a number of disciplines. These disciplines range from computer vision science, computer engineering, image processing, pattern recognition and statistics [48, 70] with FR technology being developed in two-dimensional (2D), three-dimensional (3D), Infra-red and video-based methods with the main focus being concentrated on 2D.

As the importance of FR technology has been recognised by Government departments, many resources have been committed to FR research and development. The U.S Government has funded efforts to develop FR technology and as such, the FR community has benefited greatly [8]. One of the key contributions was the development of large data sets that has initiated the development of new algorithms. In order to benchmark and test the FR performance, the U.S Government in partnership with the NIST has funded independent testing programs [1]. The programs invite universities and/or commercial vendors with capabilities in specific FR areas to participate in the testing. Funded programs include the Multiple Biometric Grand Challenge (MBGC), Face Recognition Vendor Test (FRVT) in 2000, 2002, 2006, 2012 and 2013, Multiple Biometric Evaluation (MBE) in 2010 and the Face Recognition Grand Challenge (FRGC) [8, 50, 71]. The intended aim of these programs was to evaluate technologies that could become readily operational to support the intended marketplace. The focus of the tests are dependent on the market drivers and needs of the market at that time, in which the main driver for the FRVT 2012 test is the use of FR in surveillance applications, or FIAC [72]. In particular, 'video-video, still-video, video-still' have been recognised as focal areas as "there remains considerable activity in the use of FR for surveillance applications" [72 p.9]. The most recent FRVT in 2013 set its scope to determine the level of performance in regards to recognition accuracy and computational resource usage by executing the algorithms on large databases. These databases were comprised of a mixture of reasonable quality mug shot images; moderate quality visa application images; and poor quality webcam images. The inclusion of these types of images was to show the performance rates for applications such as passport, visa and driving license duplicate detection operations, which NIST found to constitute the largest portion of the marketplace in FR. The webcam images were included to show how recognition accuracy degrades in instances where poor quality images are used, such as FIAC applications [72].

Results from these evaluations have displayed marked improvements in FR performance over the past decade, as shown in Figure 1. Results are shown at four major milestones, where each milestone show the false reject rate (FRR) with false accept rate (FAR) at 0.001 (1 in 1000). Results between 1993 and 2006 show a reported decrease in error rate (FRR) in roughly three orders-of-magnitude. These improvements can be attributed to three main developments, including improved algorithmic recognition technology, higher resolution imagery and improved quality of testing due to greater consistency of lighting [8]. A summary of the most recent FRVT 2013 shows that the largest contributing factor to the increase in recognition accuracy could be attributed to improvements in image quality. More specifically, images used for the 'known' dataset that followed, and were closest to the standards set out in the ISO/IEC 19794-5 "gold" standard produced the best, most accurate results [72].

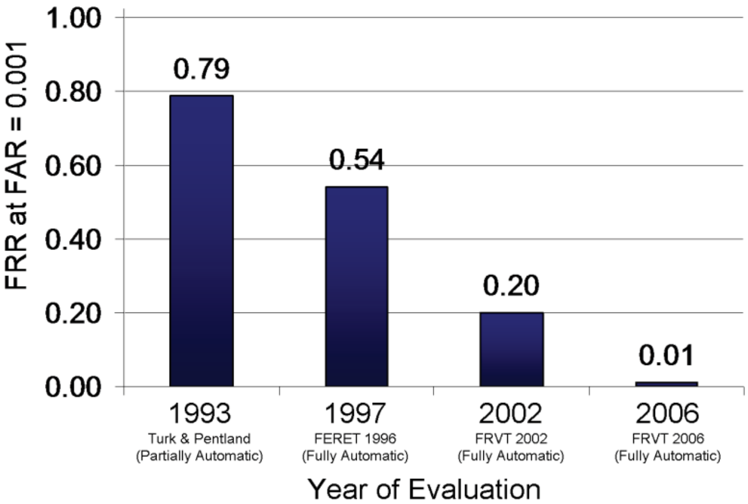

Figure 1: The reduction in FRR for the state-of-the-art face recognition algorithms as documented through the FERET, the FRVT 2002 and the FRVT 2006 evaluations. Extracted from Philips et al. 2010

With these improvements in FR performance, facial recognition technology has been implemented in both *verification* (1:1) and *identification* (1:N) scenarios [29] in which their performance varies greatly and will be discussed below.

2.3.1 1:1 Performance

Although FR has been developing since the 1960s, it was not until the 1990's that FR technology first saw a requirement for commercial applications [70]. As an example, Australian Government departments were interested in access control (*verification*) FR technology, where personnel would have their image pre-enrolled onto a database that was then used to match against their live image to verify their identity when attempting to gain access to a physical location [1]. This *verification* scenario is utilised in a controlled environment where both enrolment and live (verification) images are of high quality. The person provides their facial biometric to allow the system to compare the acquired and enrolled images. According to the ISO/IEC 19795-1 standards document, "the user makes a positive claim to an identity, features derived from the submitted sample biometric measure are compared to the enrolled template for the claimed identity, and an accept or reject decision regarding the identity claim is returned" (p.5). This is commonly used for access control scenarios. Such technology has been rolled out to a number of different applications including the SmartGate system in all Australian international airports, Auto-Gate in Brunei, UniPass in Israel and EasyPASS in Germany [63]. FR in *verification* applications is a mature field with many papers surveying performance rates, showing that the controlled environment allows for near perfect matching performance [73-75]. The 2006 Facial Recognition Vendor Test (FRVT) evaluated the top performing FR algorithms used for *verification* and concluded that the verification rate had increased from 80% in 2002 to 99% in 2006 at a false alarm rate of 0.1% [8, 75].

However, after events such as 2001 September 11 terrorist attacks, an image of suspected hijacker Mohammed Atta was retrieved from surveillance-camera footage as he passed through the Portland, MD airport metal detector [76]. It was noted in a 2001 hearing on "New Technologies in the Global War on Terrorism" by Senator Dianne Feinstein that at least two of the hijackers held accountable for the 9/11 attacks were previously known to authorities, with their photographs stored on a police database [11]. As such, and as interest amplified, the focus and direction of FR technology shifted, with attention moving from a *verification 1:1* matching application to an identification (1:N) application in uncontrolled environments.

2.3.2 1:N Performance

In order to implement the shift from 1:1 to 1:N applications, the FR system would be required to scan live faces and compare the acquired image against a stored database of images to detect wanted persons [29]. Such application would require the technology to be able to function with less constrained images taken from CCTV footage operating in an uncontrolled environment, such as the surveillance-camera image from the Portland airport depicting the suspected 9/11 terrorist. This requirement demanded the technology be able to detect a face in a crowd, compare it against a database of previously enrolled images and alert the operator of matches.

With the new requirements of surveillance and security, FR vendors began to focus on developing FR products for this purpose where competition between these rival vendors saw inflated claims as to the capabilities of FR technology, claiming near perfect performance and matching abilities [69]. After the release of the September 11 CCTV acquired image, news articles were circulated speculating that, had the correct FR technology been in place, an image like the one retrieved could have helped avert the attack [47]. FR vendors were claiming that commercially available FR technology could have 'instantly checked the image against photos of suspected terrorists on file with the FBI and other authorities' [76] in a FIAC recognition process. According to FR vendors, with the images previously on file and the acquired CCTV footage, a match could have technically been detected, potentially sounding an alarm before the suspect boarded the flight, if the technology had been installed.

However, the reality was that the FR software in 2001 was not capable of doing such a task, and now in 2014/5, the question still remains as to the feasibility of the technology for this use. Whilst FR has been shown to work well in controlled environments for 1:1 access control (such as SmartGate) [75, 77], performance in uncontrolled environments is poor [78, 79]. FR technology has reduced performance when used outside of the controlled laboratory conditions in which they are tested [79]. During a Biometric Consortium Conference in 2003, Valencia commented that:

> Performance in the lab has turned out to be a very poor indicator of performance in the field. The performance of biometrics is influenced by many factors, including the environment, user behaviour, and application [79 p.21]

Furthermore, a report released in 2009 stated that "FRT (facial recognition technology) performs rather poorly in more complex attempts to identify individuals who do not self-identify" and further, the face in a crowd scenario is "unlikely to become an operational reality for the foreseeable future" [1 p.3]. The operational performance rate of FR technology in uncontrolled environments is unknown, and to address this there is a pressing need for software to be tested in the environment in which it is intended (*in-situ*).

2.3.3 FIAC in Public Spaces

There have been several recorded attempts to trial FIAC technology in operational environments, however there is little public access to this information. Operational evaluations are often associated with privacy issues and hence the information is not freely available. There are however a few sources that discuss the use of FR in public spaces and they are as follows.

In 2001 in Tampa, Florida the U.S. Government and the local Police Department formed an initiative to trial FIAC technology in public spaces. As such, FR technology (supplied by vendor 'Viisage Technology' with software called "FaceFINDER") was implemented with a number of CCTV cameras throughout the state in a pilot project. Using FaceFINDER, a face could be acquired from the crowd in real-time and translated to a template for comparison against other stored templates (images) within a database. Viisage claimed that faceFINDER could calculate a template in both live video or digital images and search through a million stored images in a few seconds to find a match. This technology was trialled at the Super Bowl, NFL Experience and Ybor City in 2001 [81]. Results from the week-long project found 19 matches in total, but none were confirmed as no one was pursued or arrested and hence, the identities could not be confirmed [82]. None of the results of the trial were standardised or certified causing hesitation surrounding the feasibility of FIAC technology in operational environments [83]. As a result, two years later in 2003, that implementation was turned off [84].

Furthermore, in Dallas America, Visionics, another FR vendor, tested their FaceIt Argus Software in Fort Worth and Palm Beach Airport [85]. Data collected and analysed from the trial in Palm Beach Airport detailed that there were 15 participating POIs in a database of 250 people. During the trial period, the FR technology only correctly identified the POIs 47% of the time while raising 2 to 3 false alarms per hour [86]. As a result, the software was not continued as a tool for *identification*.

Asia Software, another FR vendor, have provided their software "Sova" to St. Petersburg Metro Station in Ladoga, Ministries of Internal Affairs in Kazakhstan and the Russian Moscow Metro [87]. Asia Software claims that 'Sova' can detect a persons' face and compare it against a database in a few seconds with 96% accuracy. The performance rates mentioned are released by the company and the conditions in which they were obtained are unknown. As a result, it is unclear as to whether these are the performance rates expected in an operational environment or the performance rates expected from laboratory environment and may not be suitable as a measure of the feasibility of FIAC in operational environments.

In 2007 the German federal Criminal Police Office (termed BKA) evaluated three FRS systems in Mainz rail terminal [1]. The program was a technically-oriented research project that aimed to assess the possibility of FIAC being used as a future police operational *identification* tool. In total, there were 200 participants that volunteered to create the POI database, with an average of 22,672 persons passing through the monitored area daily [88]. Results showed promise for the police with a 60% match rate with a false accept rate of just 0.1% [89]. The report revealed however, that there were factors such as lighting, camera technology and human movement that affected the performance level of FR technology. As such, the report stated that FR technology is "not yet suitable as a system for general surveillance in order to identify suspects on a watch list" [1].

With the limited information available in the public forum regarding the implementation and testing of FIAC in operational environments, it is clear that there is a need to test this application to determine whether such technology is feasible as an operational tool currently. What has become clear however is that there are a number of challenges surrounding the implementation of FIAC in uncontrolled environments that need to be addressed before trialling the technology, in order to provide the best opportunity for success. These challenges are discussed below.

2.3.4 Challenges of FIAC

In order to understand the challenges that FIAC faces, it is important to understand how FR systems typically work and how the technical, environmental and operational variables that are known to affect performance [14], present in an uncontrolled environment affect performance at each of the main technical stages of the FR process, as shown in Figure 2.

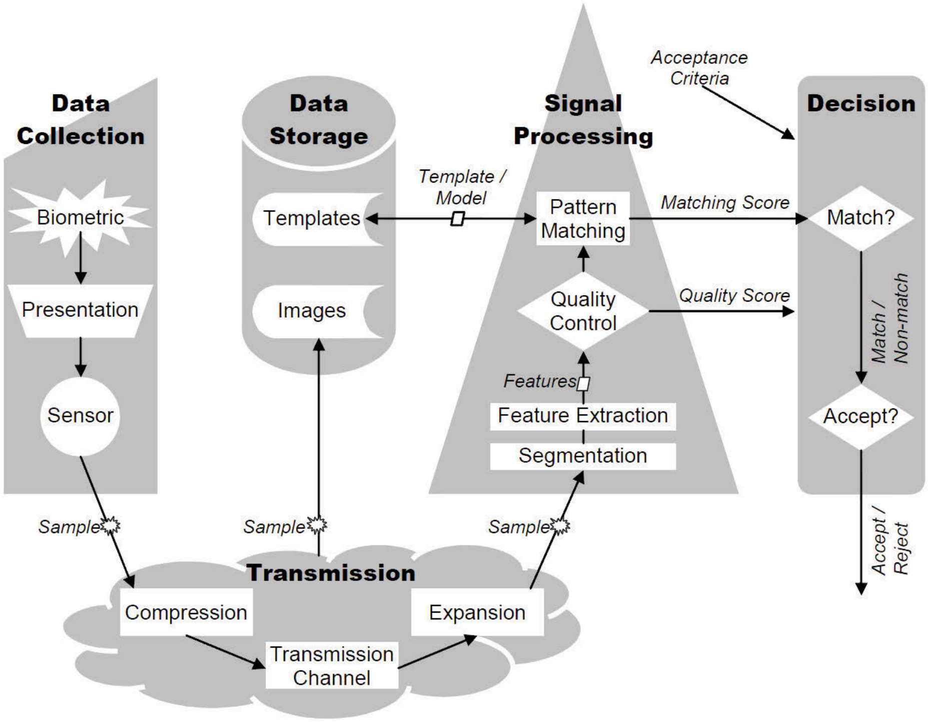

Figure 2: A Diagram of a General Biometric System as extracted from Mansfield et al. (2002)

The four main technical stages functioning at the basis of every FR system include face detection, normalisation, feature extraction and face recognition [50]. These technical stages work differently depending on whether the algorithm used is image-based or feature-based [74]. Image-based algorithms include methods such as Eigenfaces, Fisherfaces and Linear Discriminant Analysis (LDA) that attempt to identify faces using global representations [90] such as pixel intensities. These methods focus on the face as whole rather than local features. Feature-based algorithms, such as Statistical Shape Models, Active Shape Models and Active Appearance Models process the input image to identify, extract and measure distinctive facial features such as the eyes, nose and mouth [74].

There is extensive literature regarding these methods, with Rawlinson et al. (2009) providing a thorough review for reference. Both approaches create measurements of the face that are then computed into and stored as a 'vector' in a database. Vectors are then compared against other vectors stored in the database by standard statistical pattern recognition techniques to produce a match score of likeness.

Although FR algorithms function in a 'black box' manner, being that the information revealing the process is not disclosed, it is important to understand how different technical, environmental, and operational variables affect FR performance when dealing with uncontrolled environments [14]. It is very difficult to locate and extract features that are entirely immune to the variables, and so researchers have attempted to identify and overcome these factors by evaluating them [6].

Detection of a face from a facial image is not a difficult task for a human; however the same task poses difficulty for computers, as there are many variations that can affect image acquisition. For image-based models, the computer requires a mechanism to decide which pixel is included in the face, and which pixel is not [1]. For standard, passport style images this is not such a challenge, as the background is neutral and the separation from face and non-face is distinct. However, in uncontrolled FIAC scenarios, the input images are usually taken from CCTV footage in which motion, complex backgrounds and insufficient lighting are factors [91]. It is crucial that the system is able to detect a face in order to allow the downstream steps of FR to occur, where if a face is not detected, then it cannot be matched to a vector to produce a score. It is for these reasons that the conditions found to affect detection performance are kept as controlled as possible.

After detection has separated the face from the background, normalisation occurs by standardising the image in terms of size (cropped), pose (aligned) and illumination [1]. Feature extraction is then carried out to mathematically represent the face as a biometric template that is then stored in the database for future comparisons [9]. It is crucial that the correct facial features are detected, as the FR system uses these features to generate the template that then forms the basis for comparison to other facial templates. Alternatively, if the image being compared is of poor quality and the FR system cannot locate the features correctly, the system will be generating an imprecise template that will then be matched against the database. It is for this reason that the quality of the image to be matched should be controlled and standardised as much as possible.

Knowing these factors, when assessing *verification* it becomes apparent why this application of 1:1 succeeds over 1:N applications, as the FIAC scenarios are subject to conditions that impact the detection and extraction of facial images. Below is a basic description of the factors affecting FR performance in relation to 1:1 and 1:N;

- Pose: for best matching potential, images should be as close to direct frontal as possible with as many pixels between the eyes (optimal is 60 pixels interpupilary distance). In *verification* this is the case, however with FIAC, CCTV cameras are generally mounted up high and at an angle to ensure the field of view is looking down at the area suitable for surveillance. This creates pose angle problems, which are known to have a strong negative impact on FR [14] and they often operate at a distance for which obtaining the required number of pixels between the eyes is difficult [15];
- Image background: as mentioned, the FR system needs to differentiate between the background and the face being imaged. FIAC scenarios are prone to cluttered backgrounds and a large number of people in the scene passing by the cameras, which affects the direction in which the POIs are looking (being the angle of the face) as well as any objects obscuring the face such as other people, glasses, hats and/or scarves;
- Lighting: increased matching performance can be related to the lighting and illumination of the face and its features in the image. For *verification* this is often controlled however this is not typically the case for FIAC. The lighting is likely to be uncontrolled in both the POI database image and the matching (acquired) image depending on the environment (indoor/outdoor) and the time of day (day/night). Uncontrolled environments are prone to unpredictable lighting [1] with research stating "variations between the images of the same face due to illumination and viewing direction are almost always larger than the image variations due to face identity" [92]; as well as
- Expression: In *verification*, the facial expression is controlled and neutral and can be mimicked by the person seeking verification, however, with FIAC, as the images are likely acquired operationally and without the persons' direct participation from a moving crowd, the facial expressions can be varied and considering the facial features are used as points of reference for measurements, expressions are not ideal for FR matching.

These variables ultimately affect the quality of the image being produced through detection and extraction, and hence the ability of the FR system to match and identify the probe image. This brings the system to the fourth and final step referred to as facial recognition. This is the process whereby the biometric templates are compared against other templates stored in the database to generate a score of likeness in order to generate a match or non-match. However, the quality and production of the match is dependent on the aforementioned detection and extraction processes which are strongly affected by the variables present in an uncontrolled environment.

In conclusion, FIAC in uncontrolled environments presents many challenges that affect the performance and feasibility of FR technology to function in these conditions. At each stage of the FR process, there are variables that affect the success of the technology in being able to detect, extract and match a facial template that ultimately determines the overall performance of the software. Therefore there is a pressing need to evaluate this technology in an operational environment to assess the overall capabilities of the entire system including the effects of these variables in actual operational conditions.

3. Methodology

As outlined in Section 2.3 there is an overarching research problem of determining the feasibility of FIAC technology in an uncontrolled operational environment with the aim of being able to identify POIs. It has been hypothesised that each sub problem (restated below) will affect FR performance, with high resolution imagery producing the highest performance. It has also been hypothesised that FR is ready for deployment into an uncontrolled operational environment, if certain conditions are controlled for.

To examine this aim and test the hypotheses, an assessment of facial recognition performance is required, and hence several sub-questions need to be examined, which include:

1. is the level of FIAC performance acceptable for deployment in an operational environment where imagery is acquired using existing CCTV cameras? (i.e. determining operational baseline performance);
2. if not, how does the operational level of performance compare when using CCTV cameras in a more controlled laboratory environment (best possible performance using existing cameras – laboratory baseline);
3. in a controlled environment using high resolution cameras, what is the level of FIAC performance? (laboratory ceiling performance); and
4. what is the level of FIAC performance in an operational environment using high resolution cameras? (operational performance with updated cameras).

3.1 Experimental Breakdown

There are four experimental aspects of the methodology required to address the research questions, these include: an operational baseline (CCTV cameras), laboratory baseline (CCTV cameras), laboratory ceiling (high resolution cameras) and finally an operational trial (high resolution cameras).

3.1.1 Operational Baseline - CCTV

First an operational baseline analysis was conducted to determine whether FIAC was feasible in an operational environment utilising the legacy CCTV cameras. This was conducted by running CCTV footage that was previously recorded from an operational location through the latest FR algorithm. The footage incorporated all of the environmental and technical variables present in the operational environment; many of which are known to affect FR performance [14, 15]. These variables include camera resolution, field of view (angle of the face), illumination, and the number of people in the frame. This imagery once processed with the FR algorithm, produced results that would indicate the feasibility of the new FR algorithm to function in uncontrolled (operational) environments with its current capabilities and limitations.

3.1.2 Laboratory Baseline - CCTV

To reduce the impact of uncontrolled variables present in operational environments, imagery was required from a more constrained (laboratory) environment utilising the same CCTV cameras used in the aforementioned Operational Baseline above. This was done in order to determine whether the imagery being produced from CCTV cameras was of suitable quality for FR algorithms when all other variables known to affect performance were controlled. In order to test this, imagery was recorded in a laboratory setting where variables (such as lighting and angle of the face, etc.) were controlled. As such, the results produced from this imagery would be a direct indication of the best FR performance level expected when using CCTV cameras. It is expected that this would give insight into the feasibility of FR when using this type of camera technology.

3.1.3 Laboratory Ceiling – High Resolution

A ceiling analysis was also conducted in order to determine whether the performance from CCTV imagery could be improved upon by using high resolution cameras. The analysis was conducted by using imagery recorded at the same time and in the same constrained conditions as the aforementioned laboratory baseline CCTV imagery, with the only variable being the high definition GC2450 camera. As a result, the ceiling analysis (being the highest quality video imagery) would determine the likely best possible FR performance that could be expected from the current FR technology. This was examined because if results from this imagery reflect poor performance, it could be hypothesised that FR technology is not currently capable of identifying a face in a crowd and is not yet advanced enough to be deployed as an operational tool in an unconstrained environment. In this case, testing would not proceed to an operational trial. However, if this ceiling imagery produced results indicating satisfactory FR performance, then the next appropriate testing would occur in an operational setting in a location trial using high definition cameras.

3.1.4 Operational Trial – High Resolution

Lastly, the same as Section 3.1.1 an operational trial was run to determine the feasibility of FR technology to function in an uncontrolled operational environment with the only real difference being the use of upgraded camera technology. As such, a direct comparison can be made between this operational evaluation and the previous Operational Baseline results with the addition of more scenarios being included as part of this Operational Trial. Imagery was acquired from a number of locations within the environment that differed in condition ranging from 'most' to 'least' desirable and hence, different performance levels were expected. Results from these areas would allow analysis of the threshold, which is a predefined level/value, set in the FR system where values above that elicit a 'match' response, and values below do not, in varying conditions. As a result, results could indicate where the technology could and could not be utilised within the current parameters of the algorithm.

3.2 Processing and Reporting

3.2.1 Facial Recognition Algorithms

The FR testing was conducted 'offline' which means that the algorithm processed recorded (not live) footage. This was due to the fact that offline testing allows for a uniform, consistent, efficient and repeatable evaluation of the technology. It should be noted and emphasised that this is not an evaluation of a particular FR algorithm but an evaluation of the possible performance of the actual technology and no specific matcher is endorsed.

Typically, once facial identification algorithms have determined a match, a candidate list is returned to display possible identities for the probe template sample [78]. When used in an *identification* application, the threshold is typically set such that poor scores that would not meet the threshold requirements would not produce an alarm to the operator, meaning a small candidate list with fewer non-matching candidates would be returned. However, when testing FR performance as part of the research reported herein, the threshold was set so every score was returned for every possible match. This enables analysis of the core algorithmic capability of FR algorithms and the ability to determine an appropriate threshold setting. This can only be achieved by calculating all of the true and false matches including the poorest TM scores. If there is a large separation between the TM and FM scores, then determining the operating point (threshold) of the system is relatively easy in order to minimise error rates.

3.2.2 Metrics

In order to assess the performance of FR algorithms, there are a range of metrics to be used that differ depending on the stage of the FR process. FR systems are first presented with an image used for enrolment that needs to be of a certain quality in order to successfully pass through the enrolment processes and be integrated into the database as a reference template. The metrics used to assess this performance are described below. Footage then processed through the algorithm is scanned for a facial image, where, if detected, is extracted and converted into a probe template. This probe template is then compared against the reference template (previously enrolled) and a score is returned based on the algorithm's assessment of similarity. If the score (typically) is above a certain threshold set by the user, then it is seen as a match. If the images are of the same person, then it is a true match, if it is of different people, then it is a false match [93]. The scores used to assess the performance of this matching process are detailed under 'Matching Accuracy' in Section 3.2.2.2. Note however, as previously discussed in Section 2.2.1.3, for identification systems a human is always part of the FR process to make the final decision of match or no-match.

3.2.2.1 Generalised Accuracy Metrics: Failure to Acquire (FTA) and Failure to Enrol (FTE)

Both the FTA and FTE are fundamental performance metrics that are reported together and describe the ability for images to be acquired by the algorithms and converted to templates.

- The FTA is the expected proportion of transactions that fail to acquire or locate a sample of sufficient quality. This could include attempts where the biometric characteristics could not be acquired or attempts where, once a sample is detected, the segmentation or feature extraction fail due to the biometric features not meeting the quality thresholds [93, 94].
- The FTE is the proportion of samples in a population in which the FR system fails to generate repeatable templates. This could be a result of the sample (image) not being of sufficient image quality or features required at biometric enrolment, or the sample not being able to make a confirmatory match to their template in attempts to confirm the enrolment is usable [93].

3.2.2.2 Matching Accuracy: FMR and FNMR

Algorithm matching performance for score-based metrics is calculated according to the following definitions:

- False Match Rate (FMR): whereby the "proportion of zero-effort imposter attempt samples falsely declared to match the compared non-self-template" ([93 p.6]). This rate can be thought of as the proportion of people that incorrectly match to images of other people.
- False Non Match Rate (FNMR): defined as the "proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample" ([93 p.6]). Similarly, this is the proportion of people that fail to match other images of themselves.

3.2.3 Graphical Presentation of Results

- Detection Error Trade-off (DET) curves plot error rates on both axes unifying treatment of both errors, with false positives (FMR) on the x-axis and false negatives (FNMR) on the y-axis. The data is plotted on a logarithmic scale which allows separation of error rates and the ability to visually distinguish between different well-performing imagery.

- Cumulative Probability (CP) plot is used to demonstrate the relationship between the match score and the probability for both the FMR and FNMR metrics.

- Cumulative Match Characteristic (CMC) is a graphical presentation of the results of an identification test where rank (position) values shown on a candidate list are placed along the x-axis and the probability of a correct identification at that rank on the y-axis. These graphs can also give rise to information regarding the probability of false alarms presented to an operator in each of the ranks (positions) on the candidate list.

3.2.4 Statistical Analyses

In order to assess the statistical significance, it was first tested for normality using a Shaprio-Wilk (*W*) test [95]. It is important to run this test to determine whether the data meets the assumptions of a normal distribution before further analyses were conducted, to ensure the correct statistical tests (parametric or non-parametric) were being applied. The hypothesis of the Shapiro-Wilk test assumes the data would be normally distributed, so if the significant value (*p*) is greater than 0.05, then the hypothesis is accepted and data is considered normal. If it is equal to or less than 0.05 then the hypothesis is rejected and it is assumed that the data differs significantly from the normal distribution and therefore non-parametric tests are used in any further analyses.

In order to test whether there were any significant differences between two sets of non-parametric data, a Mann-Whitney U test was conducted. The Mann-Whitney U test is the non-parametric equivalent to the t-test for independent samples [96] whereby evaluations are based on the continuous variable of ranks rather than the means. The Mann-Whitney U test for two independent data sets (a and b) is calculated by first assigning all the values a rank from low to high, with the smallest number getting a rank of 1 and the largest getting a rank of n, where n is the total number of values in the two groups. U is the number of times observations in one sample (a) precede observations in the other sample (b) ranking. The U value used in the test is the smaller value from both data sets, where the lower the U, the more different the groups. To then calculate the effect size, as the sample size is greater than 20, the Mann-Whitney U Test follows the z-distribution. Results are reported according to the format set out by Morgan, S.E et al., (2002, p. 41) which require the U value, significance level (*p*) and effect size (*r*) be stated. The effect size (*r*) was calculated by the following equation whereby N = total number of cases.

$$r = \frac{z}{\sqrt{N}}$$

The effect size is a measure of relationship strength (magnitude of the results), where according to Cohen (1988 p.79), an r = .1 represents a 'small' effect size, an r = .3 is 'medium' and an r = .5 is a 'large' practical effect size [97].

Furthermore, to determine whether two sets of data were statistically related to one another, a Spearman Rank-Order Correlation was performed. This analysis calculates the strength and direction of the linear relationship between two variables. Results are reported by stating the degree of freedom (df), observed rs value and significance level (p) [95]. Descriptive statistics including the minimum, maximum, mean, median and standard deviation (StD) are also presented.

4. Operational Baseline and Determining Laboratory FIAC Performance

The operational baseline and optimisation experimentation is aimed at determining whether the quality of imagery collected by CCTV and High Resolution cameras have an affect of FR performance, in order to inform further testing and the direction of this research. Determining the relationship between camera imagery and FR performance in controlled conditions will assist in determining the level of imagery (and therefore camera type) that is required in order to provide FIAC the best opportunity for acceptable performance in uncontrolled conditions. In order to do this, provided in this chapter are methods, results and conclusions from imagery obtained in a controlled and semi-controlled environment with both CCTV and High Resolution cameras with the aim of determining whether FIAC has the potential to perform in uncontrolled environments.

This section is divided into three main categories in order to determine the core FR algorithm performance when using imagery of differing quality acquired from controlled and semi-controlled environments:

1. 2002 Operational CCTV Imagery Performance;
2. Baseline Algorithm Performance: Laboratory (CCTV) Imagery; and
3. Ceiling Algorithm Performance: Laboratory (High Resolution) Imagery.

4.1 2002 Operational CCTV Imagery Performance

The first set of imagery analysed in this section was acquired in 2001/2 as part of a previous DSTO FIAC operational location trial. Imagery was acquired from pre-existing cameras in the operational infrastructure to determine whether the resulting imagery could support a FRS. Previous analysis of this imagery using FR algorithms in 2001/2 found that FIAC was not feasible [22, 99], however the imagery was processed in the current research by a updated FR algorithm to determine the present-day performance level and feasibility of operational footage.

4.1.1 Method /Participants / Imagery type

CCTV Phase Alternating Line (PAL) cameras collected imagery from a number of locations within the operational environment for testing in 2002.

The footage chosen for the current study was obtained from an uncontrolled chokepoint scenario where the lighting was variable; however the crowd was filtered into only a few people being acquired at once.

The participants involved in the 2001/2 trial were first photographed for later use as a simulated watchlist of POIs. The POI participants were then directed to integrate into the crowd while the cameras were acquiring imagery. This allowed for the recorded footage to be used for matching between unknown people in the crowd and participating POIs on the simulated watchlist. In the current study, ninety-nine (99) participants were utilised and the initial single images were enrolled into a watchlist and the video run through the current FR algorithm.

4.1.2 Results

The histogram shown in Figure 3 illustrates the true and false match frequencies for the 2002 operational imagery with the latest algorithm. The results indicate poor FR performance as there is a large overlap of true and false match scores, with very few TM scores returning a high score. This could be due to the operational imagery being affected by variables such as lighting (both overexposure and poor lighting), pose (angle of the face to the camera), expression (facial expression), resolution and movement.

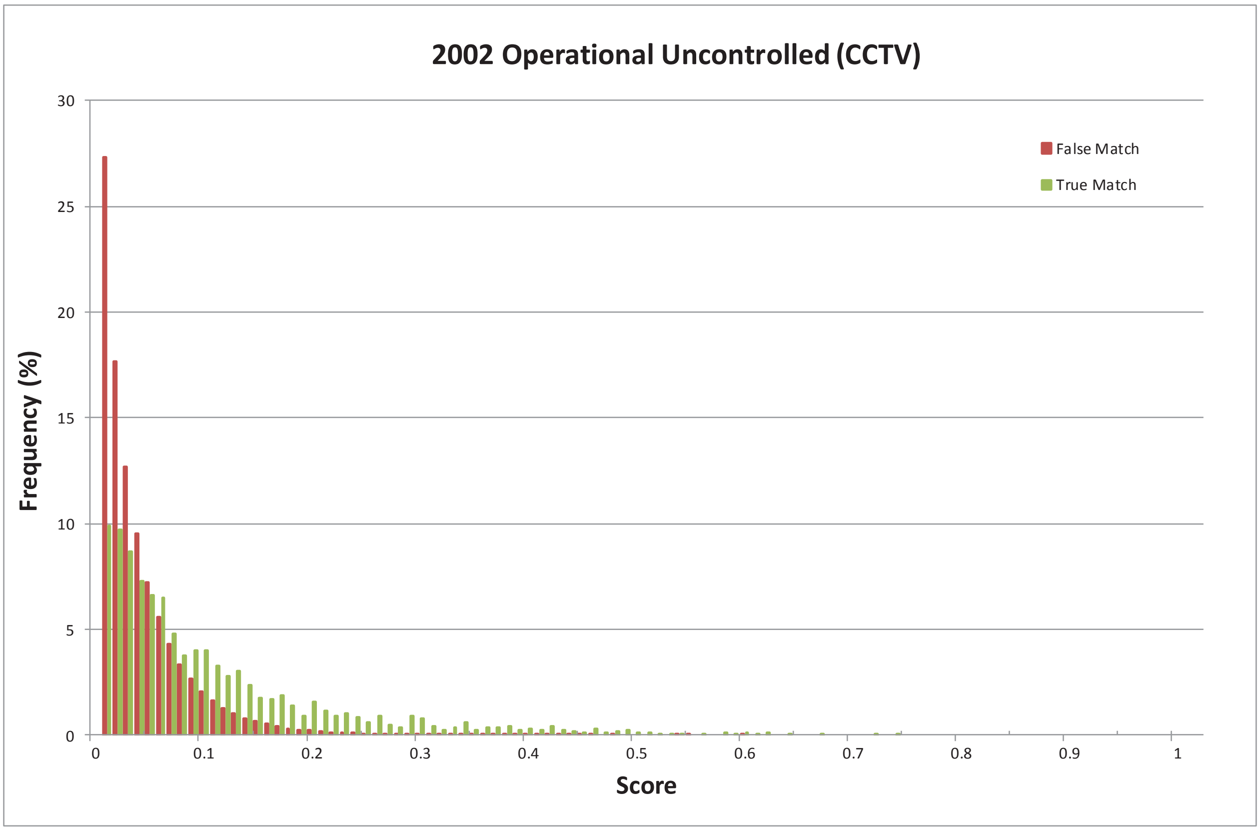

Figure 3: Histogram of 2002 Operational Uncontrolled CCTV Imagery Performance

These results suggest that the performance of FIAC when processing the CCTV operational imagery would not allow for an operating point (threshold) to be set that would produce useable results. A threshold set high would return no matches, yet a low threshold would return matches that could be true or false with a large chance of error.

Analysis indicated that further investigation was required. As such, imagery recorded in controlled settings with optimal laboratory conditions was used and is detailed in the forthcoming section.

4.2 Laboratory Imagery Performance

Facial recognition performance was tested using controlled laboratory imagery with the same CCTV cameras to assess performance when other variables known to affect FR [15, 74] were controlled. This was conducted using pre-recorded imagery from the 2009 DSTO Imaging Trial (as described in [100]). In addition to removing some of the variables that could have affected the performance of the operational footage, the same CCTV camera and a high resolution camera was also used.

The results generated from each camera (imagery) type would allow comparisons to be made based upon these quality differences. Therefore, results could potentially give insight into the possible effects that camera type has on FR matching when other variables are controlled.

To test this, the baseline (CCTV) and ceiling (high resolution) imagery were processed using a FR algorithm. For the baseline analysis, the results are shown in Section 4.2.2.1 and ceiling analysis results in Section 4.2.3.1. The technical specifications and protocol used for imagery preparation and optimisation for the laboratory testing are outlined below as well as the overarching methodology used for analysis.

4.2.1 Method

4.2.1.1 Participants

The database consisted of 314 personnel on-site at DSTO Edinburgh. Participants' images and videos were tracked throughout the imaging process to ensure the correct identity was being matched to the appropriate participant by being assigned a unique identifier and identification barcode. This was also used to de-identify the participants' images from personal information for added data security.

4.2.1.2 Imagery

The laboratory imagery used for both the baseline and ceiling analysis was collected in 2009 as part of a DSTO Imaging Trial. This imagery was acquired in constrained laboratory settings where variables known to affect FR performance were controlled and kept constant. These included illumination, resolution, field of view, pose (angle of the face) and number of people (crowd) in the frame. The imagery was collected in optimal settings that included controlled lighting, image distance and image height.

The 2009 Imaging Trial was designed so that each participant had both frontal images and video footage acquired. The still frontal "passport style" images were acquired of each participant using the Digital Single Lens Reflex (DSLR) camera, and later enrolled onto the FR system to create the simulated watchlist. The acquired video footage was recorded as the participant walked down a controlled (well illuminated) corridor towards a camera setup that included both a CCTV and high-resolution camera. Both cameras were imaging at the same time to reduce other differences (such as participant speed of movement, angle of the face, etc.). The technical specifics of the chosen camera types are shown in Table 1.

Table 1: Specifications of the Camera Types Used for Imagery Collection

Camera Type	Camera Model	Pixels (resolution)	Frame rate per second (fps)	Format	Used for
Digital Single Lens Reflex (DSLR)	Nikon D1X	1960 x 3008	Single Image	JPEG	Enrolment
Closed Circuit Television (CCTV)	AutoCam-dn	752 x 582	25 fps	PAL Video	Baseline
High Resolution	Proscilica GC2450	2488 x 2050	15 fps	GigE Video	Ceiling

The video footage of both the CCTV and high resolution cameras were saved in their proprietary format and were then converted into .avi format as a requirement for the FR algorithm.

The algorithm was configured so the settings remained constant, therefore ensuring all produced results were comparable. The ‘watchlist’ database contained all 314 participants that represented simulated POIs used for FR matching. The threshold was set so that every template being created by the algorithm was being matched to the 314 enrolled participants, and all scores were returned for analysis. Often in operational scenarios, the threshold would be set to an operating point to avoid a high number of false alarms. For testing purposes, by removing the threshold variable, it allowed a full analysis of the FR performance to be made as each template/POI comparison would return one true match and 313 false match scores.

4.2.1.3 Processing

All 314 still frontal images (acquired as detailed in Section 4.2.1.1) were enrolled into the FR database of the vendors FR software. These images created the ‘watchlist’ of POI to be compared against templates (probes) detected in the video footage (outlined below).

Video footage (CCTV and high resolution) of 100 participants[1] were chosen for the baseline and ceiling performance analysis of the algorithm. Once the chosen videos were converted from their proprietary format into the appropriate format for integration into the FR software (.avi), all videos were inputted into the algorithm "offline". This refers to the premise that the DSTO footage was pre-recorded and not 'live', which was conducted to ensure control over the FR analysis. For each facial image the FR software detected, a template was created, stored and compared against all 314 watchlist participants. Each comparison made between the generated facial template and watchlist images produced a match score that was stored in a 'journal' of results. Once the video was complete, all of the results were transferred from the journal using Structured Query Language (SQL) into a Microsoft Excel spread sheet where the data was analysed. Analysis included sorting the scores into true and false matches to be able to calculate the algorithm's performance. Scores were sorted according to rank to allow both score-based and rank-based analysis. These results are shown in the forthcoming section.

4.2.2 Baseline Algorithm Performance: Laboratory (CCTV) Imagery

The baseline analysis was conducted on laboratory controlled imagery to determine whether the quality of imagery generated by CCTV was high enough to produce satisfactory FR matching performance. As mentioned, the imagery was acquired in optimal conditions detailed in Section 4.2.1.2. The results would therefore show performance levels indicative of the best possible performance that CCTV could produce irrespective of other (uncontrolled) variables to give a clear indication of the feasibility of using CCTV cameras for FIAC applications with the current algorithm.

4.2.2.1 Results

In this section the results are presented in terms of the metrics outlined in Section 3.2.2. The histogram in Figure 4 shows the true and false match scores for the algorithm when processing the laboratory (CCTV) imagery. Results show quite low true match scores and a large proportion of the true match scores overlapping the false match scores. In Figure 4 the majority

[1] 100 participants were used instead of the 314 because of the extensive process of manually confirming each match and exporting the findings out of each algorithm.

of true matches produced scores closer to 0 than 1, suggesting that this performance is quite poor.

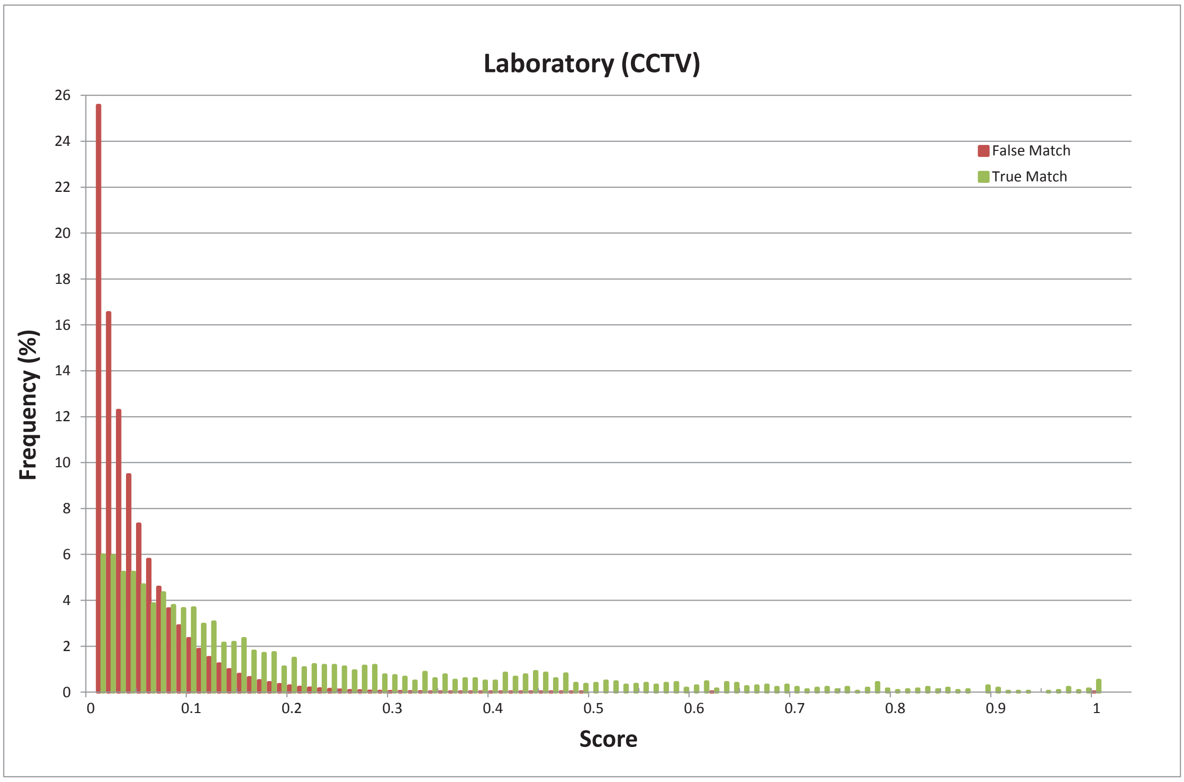

Figure 4: Histogram of Laboratory (CCTV) Imagery Performance

4.2.2.2 Conclusion

These results suggest that FIAC may not be feasible for the algorithm and imagery used. Although the imagery was controlled and optimal, the algorithm still produced poor scores with little separation between the true and false match distributions, making it difficult to potentially set a threshold that would return true matches with little frequency of error. As similar results were seen when analysing the 2002 operational chokepoint (CCTV) footage, it was concluded that the poor performance could be a result of the low resolution of the CCTV imagery, resulting in a reduced number of pixels between the eyes, which is known to effect FR performance [15]. Therefore, in an attempt to determine if results can be improved to a level that would make FIAC feasible, high resolution imagery was examined and used as a comparison to determine whether increased resolution affects (increases) FR performance prior to any further testing in an operational environment.

4.2.3 Ceiling Algorithm Performance: Laboratory (High Resolution) Imagery

The results outlined in Section 4.1.2 and further Section 4.2.2.2 reported that the CCTV imagery resulted in poor FR performance. To test whether the poor performance was a result of the camera type, high resolution imagery was processed through the FR algorithm. This high resolution "ceiling imagery" was acquired at the same time and in the same environmental conditions as the baseline imagery (detailed in Section 4.2.1.2) but with a high resolution GC2450 camera. The results would therefore show FR performance levels when processing good quality imagery from a controlled environment. These results could then be used to give an indication of the effect that the different cameras (and hence imagery) had on FR performance and be used to hypothesise the feasibility of FIAC technology (irrespective of other uncontrolled variables). That is, if FIAC is not feasible under controlled (good) conditions, with a high resolution camera, then it is unlikely it would be feasible in an operational environment where there are more variables that can degrade performance.

4.2.3.1 Results

The histogram in Figure 5 shows the true and false match scores when processing the laboratory (high resolution) imagery. Results show much-improved true match scores over that of the laboratory (CCTV) results (Figure 4) with a large separation of the true match and false match scores. This would suggest that the operating point of the system (threshold) could be set at a score that would produce a very small number of errors. Figure 5 also shows a large majority of true matches producing scores above 0.5, with the highest proportion of matches being >0.9 indicating very good performance.

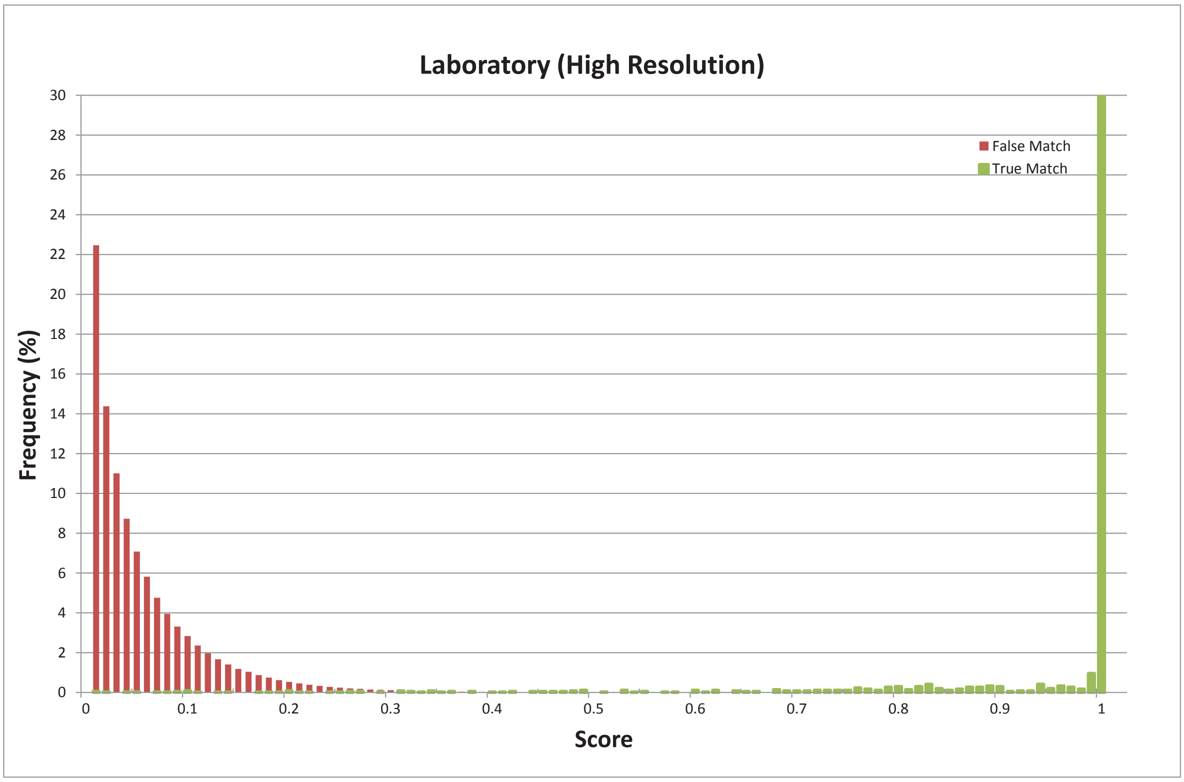

Figure 5: Histogram of Laboratory (High Resolution) Performance

In order to compare the three tested data sets to date, the overall FMR and FNMR performance is calculated and shown in a DET plot in Figure 6. It can be seen that the laboratory (high resolution) results are approximately two orders of magnitude better than both the 2002 Operational Chokepoint (CCTV) and Laboratory (CCTV) imagery at some error rates.

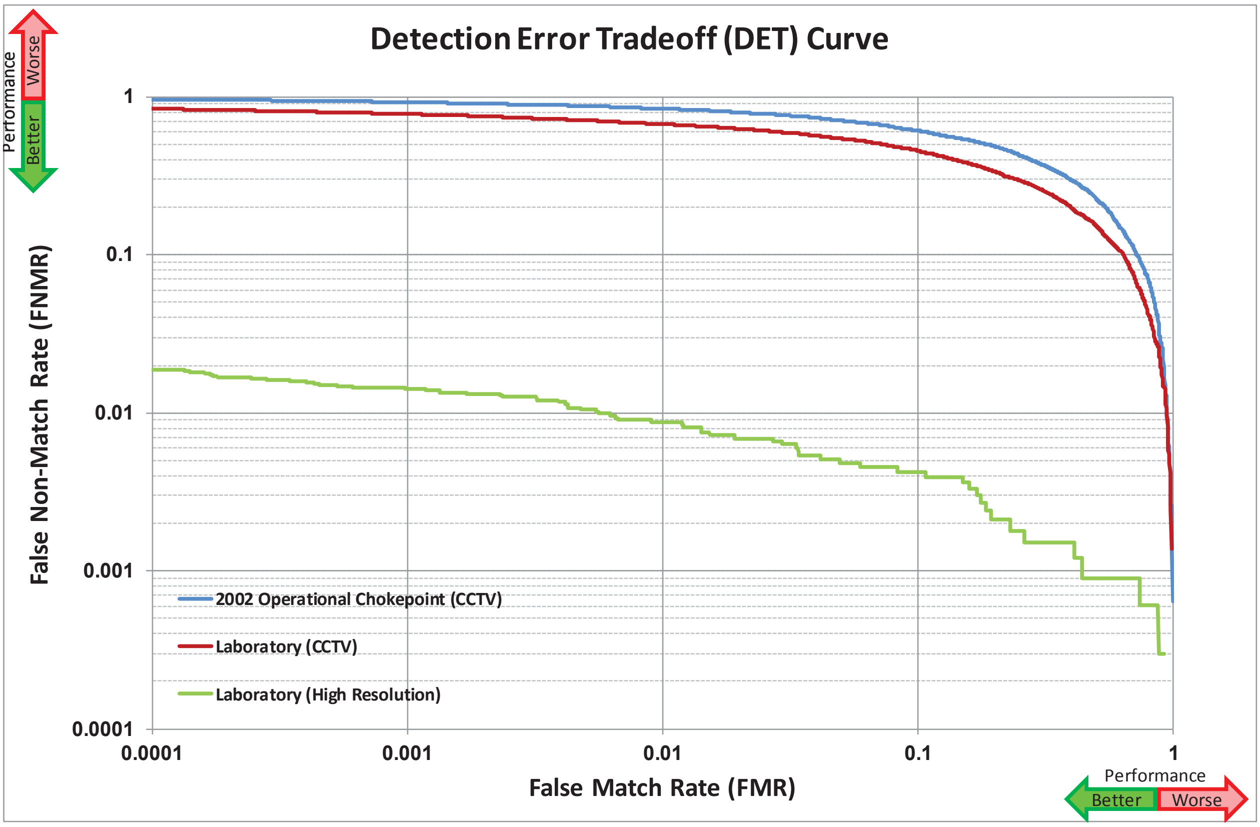

Figure 6: Detection Error Trade-off (DET) Curve Comparing the Performance of Laboratory (High Resolution), Laboratory (CCTV) and 2002 Operational Chokepoint (CCTV) Imagery.

Next, in order to show how each metric (FMR/FNMR) varied for the imagery type, a CP plot is shown in Figure 7. The primary difference to be noted is the variance in FNMR performance (how well the algorithm can later identify the same person) between the high resolution and both CCTV results. The high resolution imagery results show a FNMR that indicates a very good performance in comparison to the CCTV imagery.

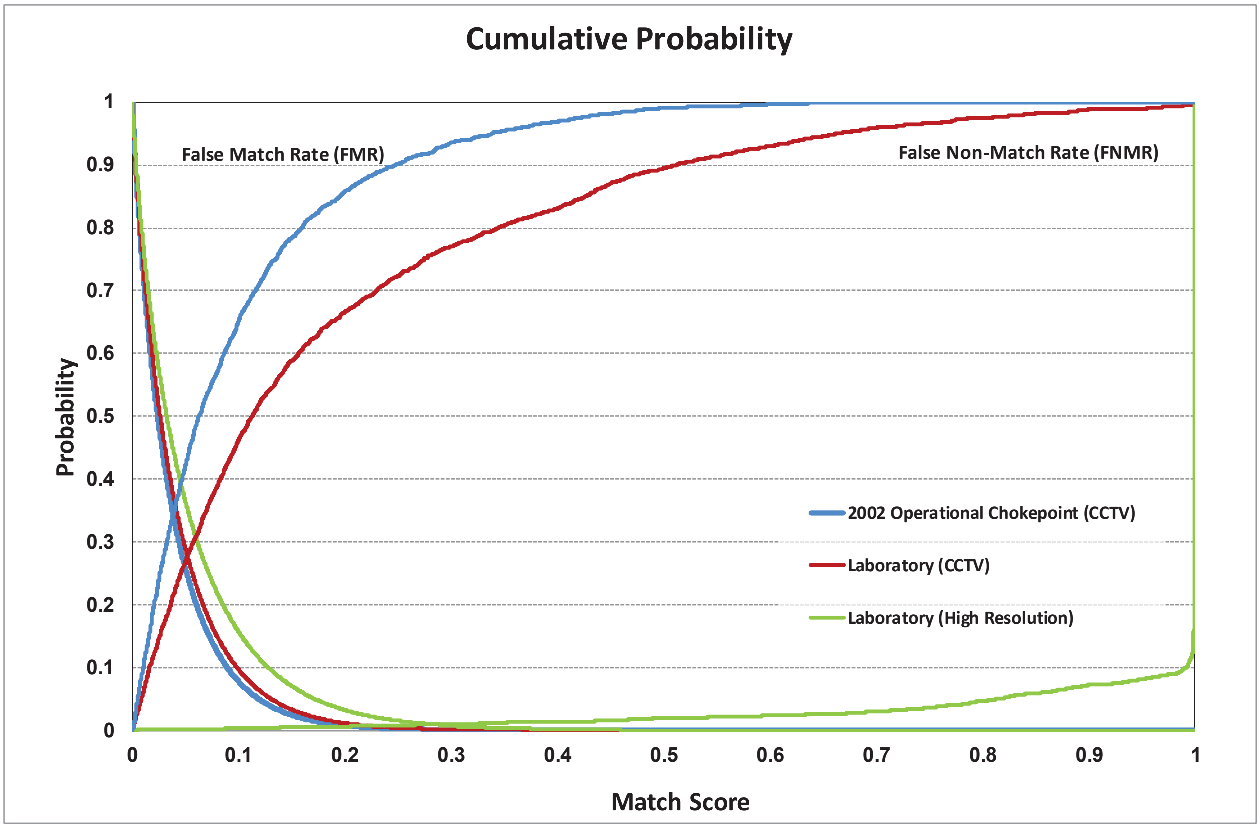

Figure 7: Cumulative Probability Plot Comparing the Performance of Laboratory (High Resolution), Laboratory (CCTV) and 2002 Operational Chokepoint (CCTV) Imagery.

In order to demonstrate the ability of the FR algorithm to correctly match a candidate to their watchlist image, a rank distribution plot is presented in Figure 8 for each imagery type. This shows the average proportion of true matches being produced in each position of a candidate list, and also gives an indication of the proportion of false alarms that an operator may need to view. The most marked difference between the data sets can be seen when comparing the ability of the algorithm to correctly place the true match in position one of the candidate list. The laboratory (high resolution) imagery produced a large proportion (94.9%) of true matches in position one compared to 20.1% for the 2002 operational chokepoint (CCTV) imagery and 25.9% for the laboratory (CCTV) imagery.

Further performance differences can be seen when comparing the cumulative frequency of true matches that occur in the top ten positions, with the laboratory (high resolution) imagery producing true matches in the top ten positions 95.8% of the time, compared to 49.1% for the 2002 operational chokepoint (CCTV) imagery and 41.6% for the laboratory (CCTV) results.

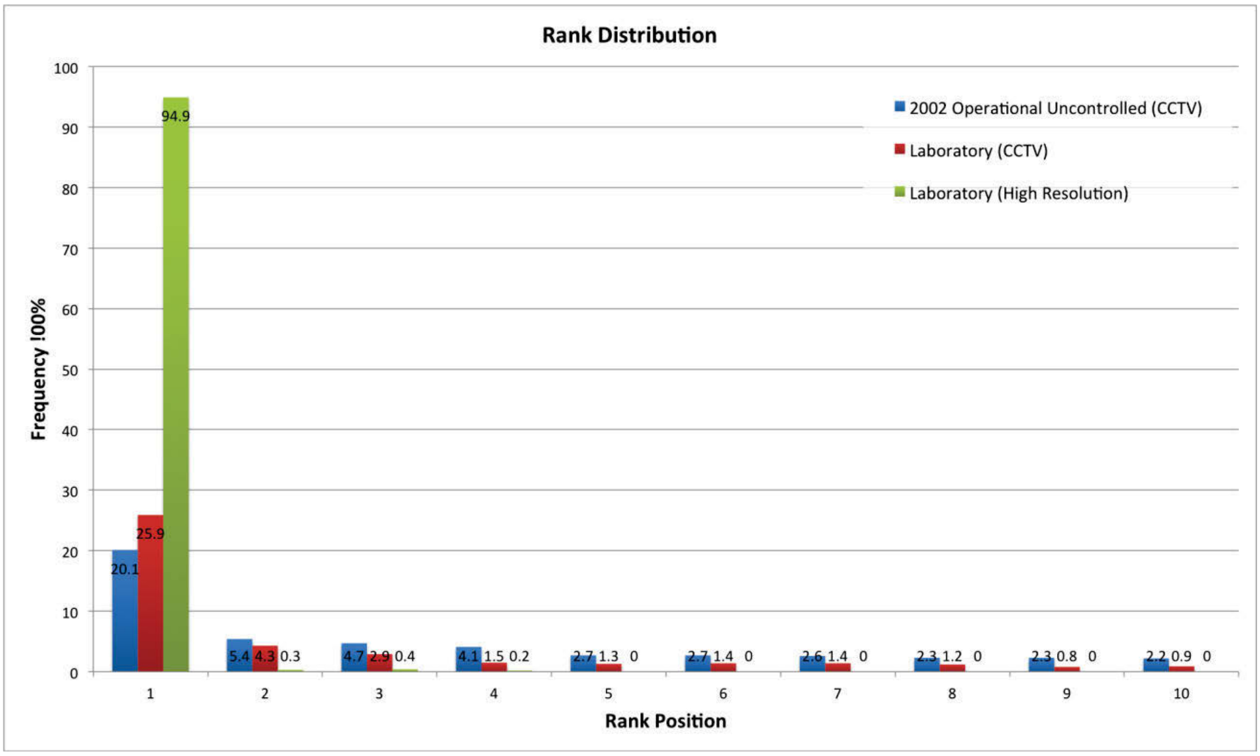

Figure 8: Cumulative Match Characteristic Comparing the Performance of Laboratory (High Resolution), Laboratory (CCTV) and 2002 Operational Chokepoint (CCTV) Imagery

As the results show performance differences between the two camera types when used in the laboratory environment (where nothing else was varied), it was concluded that the primary reason for this was likely due to resolution and hence the number of pixels between the eyes (CCTV 752 x 582 and the high resolution 2488 x 2050). In order to test this hypothesis, the laboratory imagery for both camera types were further analysed (therefore excluding the 2002 operational imagery) to calculate resolution differences without any other variable changes.

In order to examine this further, the true match scores (which was the primary metric that appeared to be limiting performance) were examined with respect to resolution (the number of pixels between the eyes) and this is shown in Figure 9. Whilst this is not a common method to measure FR performance, it is hypothesised in this research that pixels, resolution and FR performance are related. Therefore, in order to calculate this, for every true match acquired, the number of pixels between the eyes was extracted from the results (which was a metric recorded by the FR algorithm). Once all of the true matches had a related pixel measurement, the two metrics were examined with relation to one another to determine if there was any correlation between the two. It can be seen that generally, as expected, the laboratory (high resolution) results showed typically higher true match scores than the laboratory (CCTV) results.

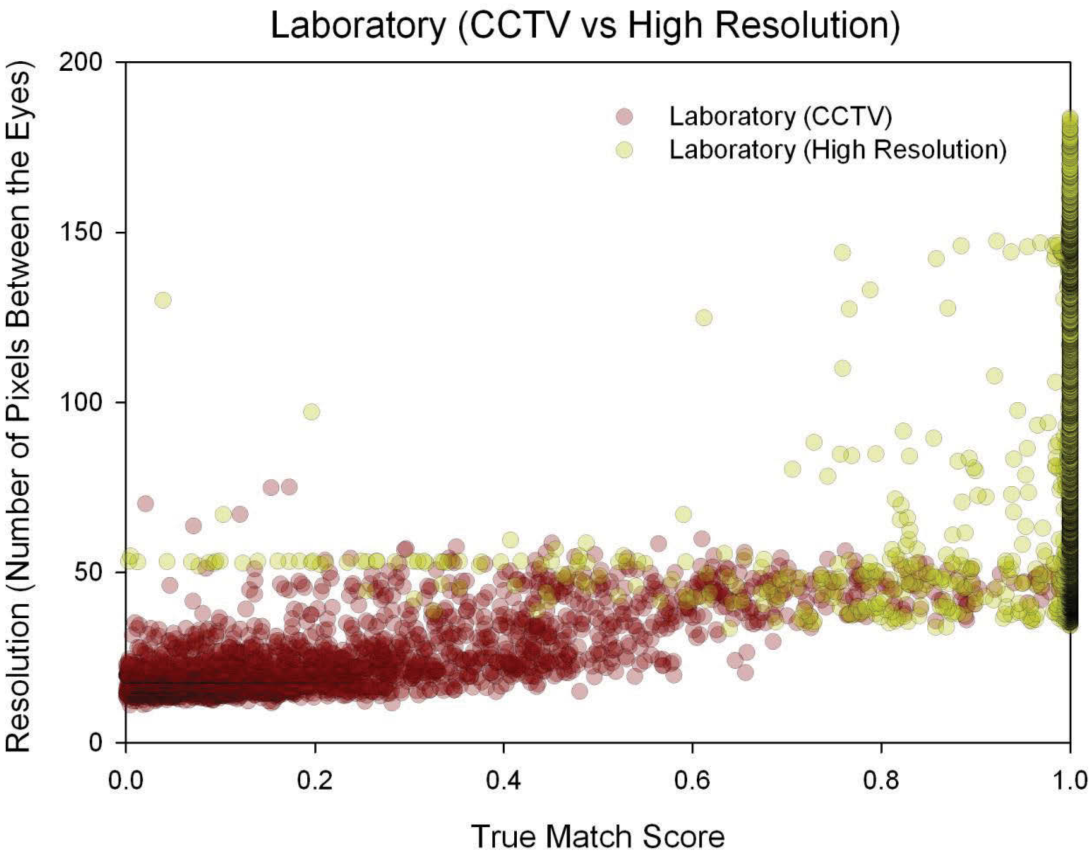

Figure 9: Distribution Scatter Plot Comparing CCTV and High Resolution Match Score and Corresponding Pixel/Resolution

These findings suggest there is a relationship between camera resolution, pixels between the eyes and matching performance of the FR algorithm. To determine whether this observation is statistically significant and affecting performance considerably, the TM scores and number of pixels were tested for normality using the Shapiro-Wilks test. The results produced significant differences (p=<0.05) for both score and pixel when examined independently. Hence the distribution for both data sets could not be considered normal and as such, non-parametric statistics were applied for further testing. Descriptive statistics were then generated and are shown in Table 2.

Table 2: Descriptive Statistics of Score Results and Pixels between the Eyes for Laboratory (CCTV) and Laboratory (High Resolution) Imagery

Imagery Type	N	Mean	SD	Median	Min	Max
Pixel Based Analysis						
CCTV	2900	23.62814	9.99647	19.96969	11.1300	75.1481
High Resolution	3292	77.13249	40.17236	59.46743	33.5786	224.2796
Score Based Analysis						
CCTV	2900	0.19838	0.21327	0.11379	0.00020	0.99950
High Resolution	3292	0.96880	0.12044	0.99952	0.00249	0.99960

A Mann-Whitney *U* test was conducted to analyse if there were significant differences between the number of pixels and scores produced by CCTV and High Resolution imagery respectively. Firstly, as expected, results from the **number of pixels** between the eyes showed a significant difference (U= 294355.0, p=<0.01, r= -0.82) between CCTV imagery (n=2900) and the high resolution imagery (n=3292). The results from the **score data** also showed the CCTV (n=2900) and high resolution (n=3292) were statistically different (U= 79306.0, p=<0.01 r= -0.85) with high resolution producing significantly better scores.

Further analysis was then conducted to determine whether there was a correlation between the number of pixels between the eyes and the subsequent TM score. The analysis was conducted using Spearman rank-order correlation (r_s) and found that CCTV (r_s(2898) = .573, p=<0.01) and High Resolution (r_s(3290) = .276, p=<0.01) scores were both positively correlated with the number of pixels between the eyes, where an increase in one is paired with an increase in the other.

The analysis was then further explored to determine whether there was a general trend in the relationship between the number of pixels and the TM score irrespective of camera type. Results found a positive significant correlation between the two variables (pixel and TM score) (r(6190)= .686, p=<0.01). These results suggest that for the camera types tested, when there is an increase in the number of pixels between the eyes, there is also an associated increase in the true match scores.

Therefore it can be concluded that the high resolution camera produced imagery that has an additional number of pixels between the eyes of the participants which also saw a significant increase in TM scores and as a result, increased performance.

4.2.3.2 Conclusion

In conclusion, controlled Laboratory 'baseline' (CCTV) and 'ceiling' (High Resolution) imagery were processed through the FR software to determine whether the current FR performance was advanced enough for potential deployment into an operational environment. The results showed that the High Resolution imagery produced much better FR performance when compared to CCTV imagery.

Whilst this is encouraging, it is not possible to extrapolate these findings to determine if these results provide an indication of the potential success of FIAC in an operational environment. This is due to the fact that the tested imagery was acquired from a laboratory environment where variables present in real-life scenarios that are known to affect FR performance were controlled, and therefore not a contributing factor in the imagery. Hence, to determine if FIAC is feasible in an operational environment, an operational evaluation is required whereby imagery is collected using high resolution cameras.

5. Operational Evaluation

The current research results have shown that the tested FR algorithm functions at an acceptable performance level when the imagery input is recorded using high resolution cameras in controlled conditions. However, the imagery tested to date was acquired in optimal conditions with the POIs actively participating in the process. In order to determine the performance of FR for FIAC applications, it is necessary to test imagery acquired from uncontrolled locations to indicate performance of FIAC in operational environments.

However, there is rarely one single consistent environment that could be used to define the operational FIAC scenario. Hence, in order to test and analyse the effects that different environments have on FR performance, imagery from three indoor operational locations was acquired that had various imaging conditions. These can be defined as controlled, semi-controlled and uncontrolled for examination purposes. The environments in which the imagery was recorded ranged from narrow 'choke points' where the crowd was imaged in single file to wide-angle open areas where there was a larger number of people present in the field of view. Results are shown in terms imaging scenario, starting from the best likely performance to the least likely performance, and were tested in a deliberate and methodical manner to enable a thorough understanding of where the FR performance is likely to work best. The scenarios are as follows:

1. 2012 Operational Controlled
2. 2012 Operational Semi-Controlled
3. 2012 Operational Uncontrolled

5.1 Method

5.1.1 Participants

Imagery was collected from high resolution cameras from three locations within the operational environment as defined in Section 5.1.2.

The trial participants comprised of enrolled and unenrolled individuals. The watchlist used for experimentation was comprised of passport-style, high quality images of POIs that were unaware of the imaging process. For the purpose of the operational evaluation, temporary watchlists were created for each scenario including a standard false match watchlist which was comprised of 200 participants and used for every location. The false match watchlist was created to maintain consistency as this watchlist contained no positive matches and could be used consistently across all locations. The true match watchlist varied in number for each location, and were as follows:

1. 2012 Operational Controlled: 129 participants
2. 2012 Operational Semi-Controlled: 234 participants
3. 2012 Operational Uncontrolled: 99 participants

5.1.2 Imaging Scenarios

5.1.2.1 2012 Operational Controlled

This scenario/location presented the most optimal conditions of the three operational environments, giving FIAC potentially the best opportunity to function and acquire faces. The lighting was ideal for FR with uniform, diffuse illumination directed straight upon the face, eliminating any shadowing created by other lighting forms such as downlights. The crowd was guided by a snake-line into single file allowing the camera optimal opportunity to image one face at a time. As such, the acquired face consumed a large percentage of the image meaning there are a large number of pixels between the eyes. These conditions are optimal and not typically defined as FIAC, however the environment and location did allow for a baseline assessment that will indicate the best possible operational performance obtainable.

5.1.2.2 2012 Operational Semi-Controlled

This semi-controlled location had the camera positioned in an indoor area that was exposed to a number of environmental factors that differ in FR suitability. The main source of lighting in this scenario was provided by down lights in the ceiling, which varied in intensity and could be attributed to causing shadowing on the face. This operational location had a larger number of people passing by the camera, and although moving slowly, the participants were not constrained in number or where they looked. This therefore required the camera to have a wider field of view to acquire all faces and as a result, an acquired face consumed a smaller portion of the image resulting in fewer pixels between the eyes.

5.1.2.3 2012 Operational Uncontrolled

This operational environment presented the most difficult scenario for FR to perform as the most number of uncontrolled variables affected this area. These include the area being wide and open and often having a large number of people passing by at any one time. As such, the camera field of view was set wide, ensuring all passing people were potentially able to be detected and have their face acquired. As a result, the facial images were small and had fewer pixels between the eyes. This environment was also affected by varied lighting as the area contained windows that subjected the imaging area to the changing outdoor environmental lighting which affected the quality of the imagery i.e. potential overexposure during daylight and dim lighting during dusk. The location was also an area where people were moving relatively fast. When the imagery was being analysed, it was observed that faces acquired further away could be obscured by those people in front of them making it difficult to acquire a good quality facial image of all people.

5.1.3 Processing

In order to analyse the performance of FIAC, operational footage and a *watchlist* were required. The *watchlist* is typically a database of POI images that an operator is searching for. The reason for this can vary as mentioned in Section 1, however for the current research, these watchlists were used to determine the true and false match rates of the FR technology.

The processing involved the following:

1. create watchlist and enrol images into the FRS;
2. initiate the operational footage from one location in FRS;
3. export the results from MSSQL into excel;
4. translate and export the generated transaction ID images from the FRS journal using Python scripts;
5. group the transaction ID images into each individual (could contain 100+ images per person);
6. manually match the POI watchlist image against the transaction ID individual;
7. record which transaction ID individual were true matches and which were false matches; and
8. pre-process the data for analysis.

As there was no ground-truthing information for aligning the watchlist images to the transaction ID images, the manual process of matching these was very time consuming. In the Operational Uncontrolled scenario, over 68,000 transaction ID images were generated that needed to be grouped into individuals to be identified as either true or false matches. This large number could be due to the fact the FR threshold settings were set to zero for testing purposes to enable a full analysis of the capabilities of the system. However, in an operational scenario, a threshold (operating point) would be set so that an image that returns a match score below a certain point would not be returned as an alert to the operator. Ideally, in an operational system, if there was a match against the watchlist that generated a score above the threshold, only one image would be returned as an alert for the operator to confirm (known as tracking). However, the current system utilised in this research did not currently recognise the same person from frame to frame, therefore often producing hundreds of images per person that were all treated independently. Newer algorithms have developed and evolved to include the function of 'tracking', however at the time of analysis the current algorithm did not contain this ability.

5.2 2012 Operational Trial

5.2.1 Results

The histograms shown in Figure 10 illustrate the true and false match scores for the imagery produced in the three operational scenarios (controlled, semi-controlled and uncontrolled) similarly to that in Section 4.2.3.1. The main point of focus for these histograms is the area between true and false match scores, with the ultimate aim being to have a large area of separation in which a threshold would be set. It can be seen that the FMR is relatively similar between all three locations, with the largest performance variability being the true matches.

The ability of the FR system to correctly identify true matches appears to increase as environmental control increases, with the controlled imagery showing the least amount of overlap between scores, followed by the semi-controlled and uncontrolled. The poorer results indicate that it would be difficult to set a threshold that would return an adequate number of true matches with certainty, as the overlap between true and false matches cannot be separated.

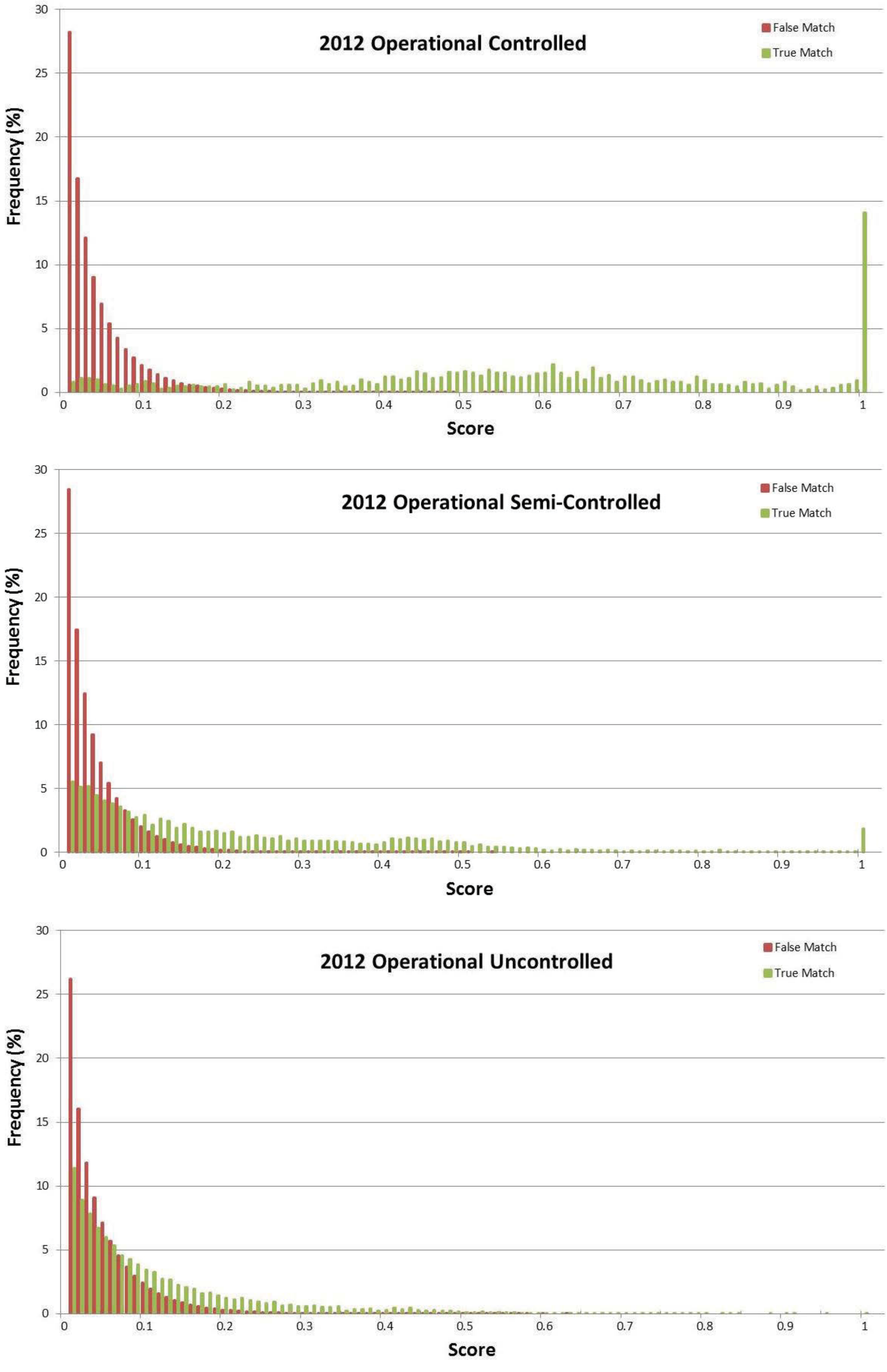

Figure 10: Histogram of Operational High Resolution Imagery from the differing Locations

The results were then expressed in a DET curve shown in Figure 11 to illustrate the overall FMR and FNMR for the operational high resolution imagery. The 2002 operational imagery, and both CCTV and high resolution laboratory imagery have been included to give context to the 2012 results and show any performance differences.

Results indicate the best performance for the most controlled environment, with performance dropping as the locations became less controlled.

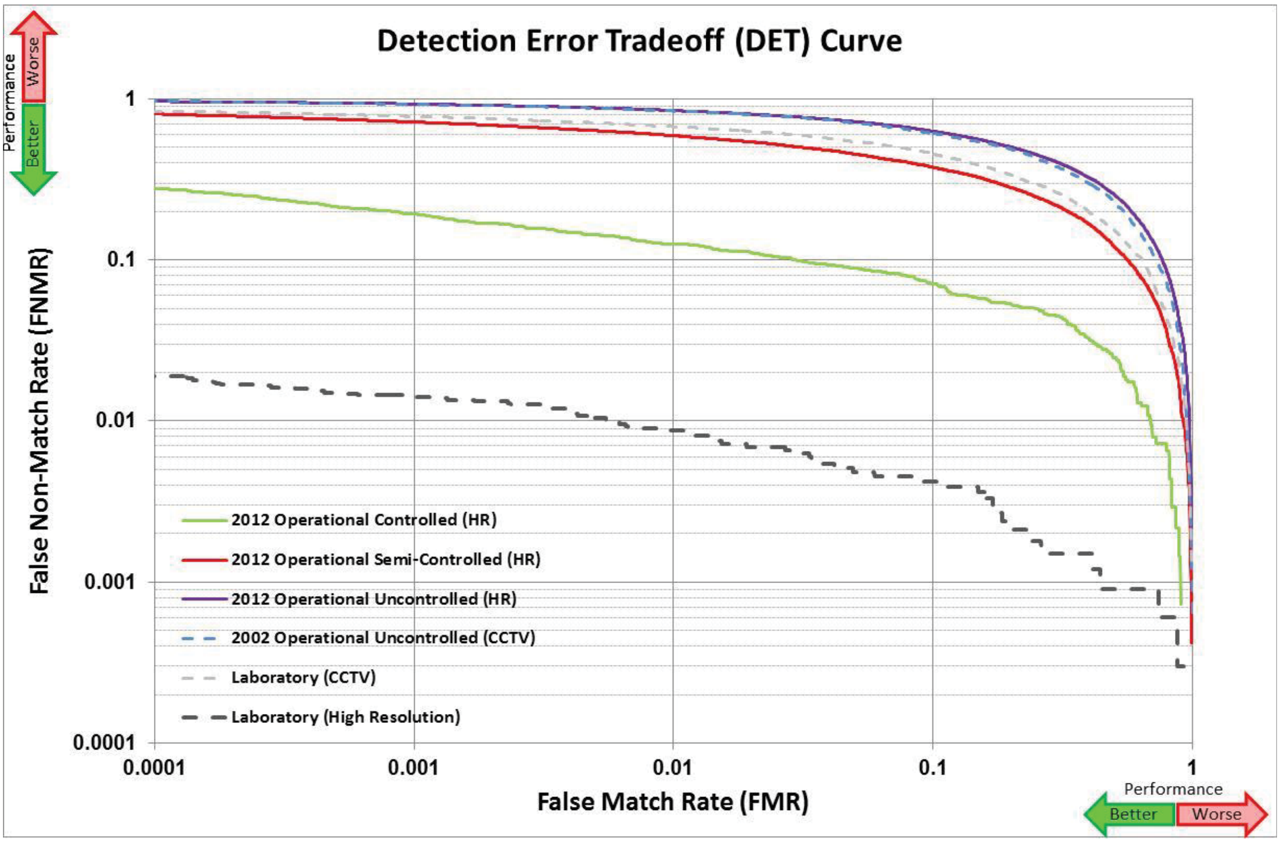

Figure 11: Detection Error Trade-off (DET) Curve Comparing the Performance of the 2002 Operational Uncontrolled (CCTV), Laboratory (CCTV and High Resolution) and the 2012 Operational Imagery

It is encouraging to note that the imagery acquired from the 2012 operational controlled and semi-controlled environments performed better that the laboratory CCTV imagery that was acquired in optimal conditions. This indicates that the imagery acquired from a less constrained environment from a high resolution (quality) camera can produce imagery that yields improved FR matching than imagery acquired in the most constrained environment.

Another point to note is the seemingly inseparable performance rate of the 2012 and 2002 uncontrolled imagery sets. This indicates that the addition of a high resolution camera in 2012 does not appear to affect the FR performance from imagery recorded in this environment, suggesting that there are other factors affecting FR matching performance. Yet, it is important to note that although both uncontrolled locations were the same imaging scenario, just ten years apart, there were other factors that were not constant between the two image sets including camera angle, lighting and the age differences between enrolment (watchlist) image and participant in the crowd. It can therefore be hypothesised that camera resolution (and therefore number of pixels between the eyes) may not be the only variable affecting the uncontrolled imagery to return poor performance levels.

Next, the CP plot shown in Figure 12 further illustrates the performance differences between the three 2012 operational locations compared to the 2002 operational imagery and the laboratory CCTV and high resolution imagery.

The primary trend to note within the 2012 operational imagery is the large disparity in FNMR performance, as the FMR has little variability and remains relatively constant between all data sets. Again, it can be seen that the controlled operational environment produced imagery that performed best, with a reduced probability of returning a low score compared to the imagery collected from the other two operational locations.

Although the 2012 controlled imagery does not perform as well as the high resolution laboratory imagery, again it is encouraging to note the increased performance levels of the controlled and semi-controlled 2012 operational data compared to the CCTV laboratory imagery.

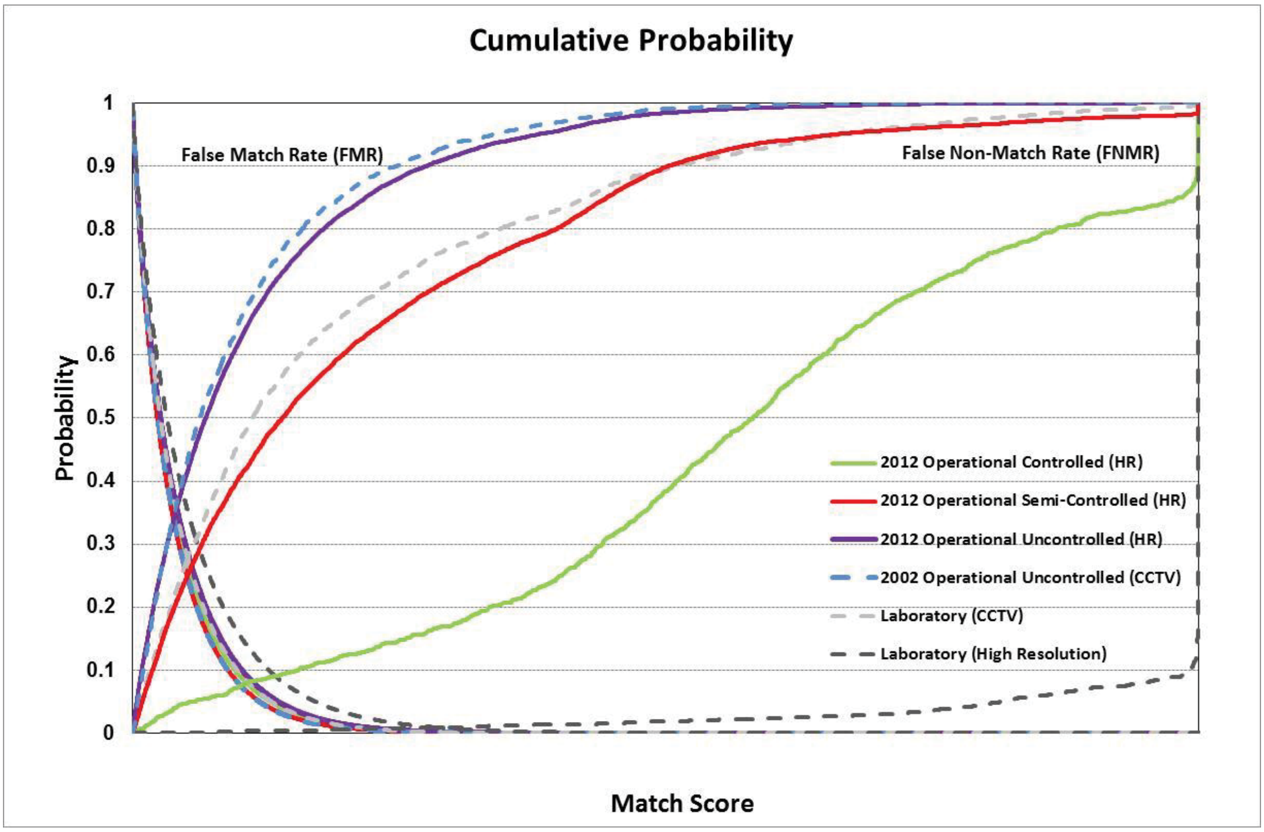

Figure 12: Cumulative Match Characteristic Comparing the Performance of the 2002 Operational Uncontrolled (CCTV), Laboratory (CCTV and High Resolution) and 2012 Operational Imagery

In order to understand the performance of each imagery set in an operational context, the results are displayed in terms of the rank data, displayed in Figure 13. This representation illustrates the ability of the FR system to acquire an image that could be correctly matched to the true watchlist candidate (TM) from each operational location, and that returned in a candidate list. These results are particularly important for operational environments, as these are the matches that would be presented to an operator. The controlled imagery would return the candidate in position one, on average, 84.6% of the time, compared to 48.2% for the semi-controlled and 10.9% for the uncontrolled. However, as an operator would not just look at the first returned candidate for a potential match (but likely the top 10 candidate returns), the performance levels for the true match being returned in the top 10 positions is 91.2%, 64.2% and 27.1% respectively.

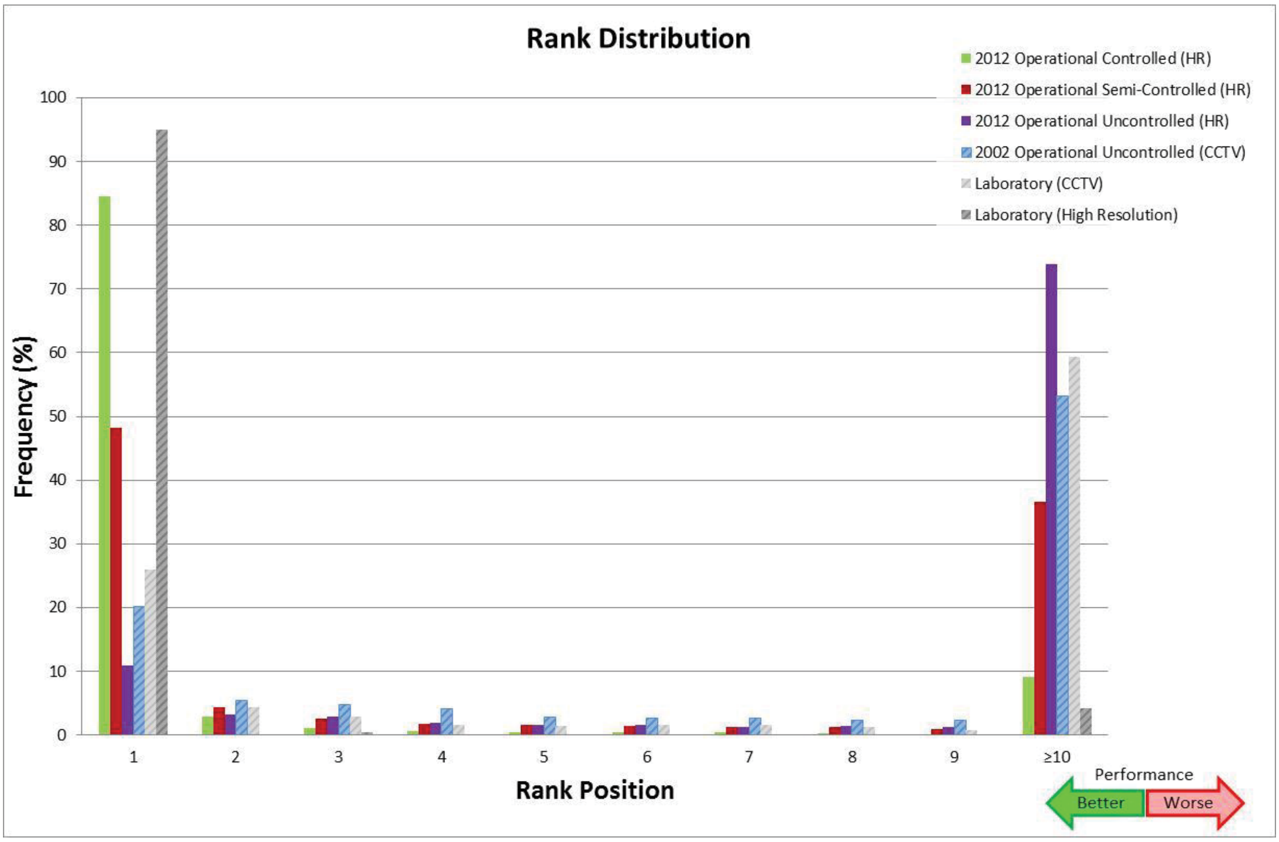

Figure 13: Rank Distribution Histogram Comparing the Performance of the 2002 Operational Uncontrolled (CCTV), Laboratory (CCTV and High Resolution) and 2012 Operational Imagery

5.2.2 Conclusion

The aim of the operational experimentation was to determine the feasibility of FIAC technology in a semi-controlled environment. This cannot be answered with a simple explanation, as there are many technical, environmental and operation variables that affect the feasibility assessment of this technology. Based on the results produced from the 2012 Operational Evaluation imagery, the trends show an increase in performance from previously acquired 2002 operational footage. The results suggest that the greater the control over the environment, the greater the possibility of FIAC being able to operate effectively. If the environment in which the cameras were to operate were controlled as much as possible, then FIAC may be feasible.

Results showed that, as expected, the controlled location produced the most promising performance, where the participants were looking straight ahead, lighting was uniform (optimal), there was only one person in the frame at a time and they were stationary. Although it could be argued that this environment was not truly a FIAC application, the imagery produced does show the best possible performance that could be acquired in an operational location with this FR algorithm.

The other two environments (semi-controlled and uncontrolled) show reduced performance and facial matching ability in comparison. This could most likely be attributed to the cameras imaging participants off-axis, therefore producing non-frontal images (making matching more difficult), the non-uniform lighting in the location and/or the necessity for the camera to be set to 'wide' to ensure all participants were imaged, even though that meant a decreased number of pixels between the eyes.

During the ground-truthing of this data, a number of observations were made as to the possible reason performance levels were low in some instances. It became apparent that in almost every instance a face was visible in the camera field of view; the algorithm could detect, acquire and match that face. However, there were many occurrences in the semi-controlled and uncontrolled environments where faces were obscured by a person standing in front of them. Although when this person moved, the person standing behind was detected, their exposure to the camera was brief, resulting in only a minimal number of transaction ID images (that were of poor quality). This then affected the matching ability for that participant, lowering the overall performance for that operational location.

Other observations that affected the FR algorithms ability to detect and match faces was the lack of good quality full-frontal images obtained from the semi-controlled and uncontrolled locations. The participants in these areas were distracted by the happenings' within the environment meaning they were constantly looking around making it difficult to acquire a full-frontal image. Some participants were almost entirely missed by the camera as they were either looking down at the ground as they walked past the camera field of view, were on the phone or glancing out the window. This aspect, accompanied by poor lighting meant that in the cases where a frontal image was acquired, the quality was poor as it was affected by shadowing, angle and poor resolution. The resolution in these areas, as mentioned previously, was poor as the camera was set to a wide setting to ensure all participants moving through the area were acquired. However, in doing this, the number of pixels formulating each face was decreased, therefore reducing resolution and the matching capacity. It became clear also, that the algorithm was detecting faces in the background and attempting to match them to the watchlist. Although this shows a marked ability of the algorithm to detect faces, as there was no threshold set for the testing, the resultant match scores were very low. This affected the overall TM and FM scores that could have been excluded had a minimum score threshold been set therefore increasing the overall performance. This could be done through tuning the FR system to ensure faces in the background (that are too small) are not acquired by the system.

The other important observation to note that affected matching performance, was participants wearing obscuring clothing such as sunglasses and hats. It appeared as though the algorithm could detect the face easily which is an encouraging improvement, however the matching scores were low and often false (FM).

The FR algorithm detected a face in every frame independently. That is, there was no tracking of people from frame to frame. This resulted in a large amount of matches (and uses a lot of computational power) for one person and uses every image it detects independently, even if it is of poor quality. As of the time this research was being undertaken, some FR vendors claimed their systems had the new capacity to 'track' a person, however this ability was not tested. It would likely be beneficial if a FR algorithm could track a person and create a single template for each person. This would result in less (predominantly false) matches being made.

Overall, it was promising to see that the FNMR performance for the controlled and semi-controlled locations have improved from the 2002 operational imagery, showing that the algorithms, computer processing power and increased resolution of cameras has increased performance potential.

6. Conclusions and Future Directions

For many decades, there has been a requirement to identify unknown persons. In the 21st century, with technology rapidly developing and the introduction of biometrics, new methods of identification have been established. However, moments in recent history have highlighted the areas of national and global security that need further development to prevent potential security problems from reoccurring in the future. This can especially be applied to assisting in identifying POIs that security applications dictate require detection. In cases such as these, current means of automated biometric identification that require participation such as fingerprint technology cannot be utilised. A form of technology was required that allowed surveillance and identification that was contact and participation free. It is for this reason that facial recognition, and in particular "face in a crowd" facial recognition needs be developed to be able to detect, compare and match a facial image from public spaces to a database of POI images. This could then be applied to scenarios such as detecting known offenders, missing persons, persons attempting to flee a country unlawfully and/or wanted persons using an alias or fake identity.

Currently, research has shown that in FR *verification* (1:1 scenarios), the combination of automated FR technology and human adjudication has acceptable performance rates. For *identification* applications however, especially in FIAC scenarios, there is limited available operational research. Therefore, there is little access to data to provide 'real life' performance measures meaning there is a gap in the current research in providing information to interested organisations for potential deployment of FIAC into locations.

In a means to overcome the lack of publically available performance data, DSTO undertook a number of research projects to assess the performance levels of FR systems for use in FIAC applications. Results from the 2001/02 operational trial concluded that the performance ability of FR algorithms was poor and not developed enough, and hence an operational capability was not possible at that time. Therefore, in 2006, another operational evaluation was conducted that revealed algorithmic improvement in FR matching performance, however the computing processing power was too slow to host FR matches in a time-sensitive operational environment. Further to this in 2009, imagery was collected in a laboratory and used to determine base performance rates using a 2012 algorithm which found that CCTV imagery produced poor FR performance rates while HR imagery had near perfect matching performance. This gave the basis that the FR matching technology was developed enough when presented with high quality imagery. As this imagery was obtained from a controlled laboratory setting, it was then necessary to evaluate the technology using high resolution cameras in an uncontrolled environment, which was conducted in the current research 2012 Operational Trial.

The results of the 2012 Operational Trial revealed increased computing power from testing in 2001/02, and improved FR algorithm speed and accuracy, namely face detection and matching improvement. There are still, however, many areas that need further development before unconstrained operational deployment would be feasible. The research initially indicated that performance in controlled environments is good, considering that it was a real test and used images that were affected by time and age. As the current research evaluated three different scenarios with three differing levels of environmental variability (control), the results gave insight into the threshold of environmental control needed to produce imagery that was conducive to accurate and acceptable matching results. The research highlighted as hypothesised, that the scenarios that produced good FR performance had the most environmental controls. This suggests that if the locations in which the camera was going to be placed had similar environmental controls, and these testing conditions were simulated in the other locations, similar results could be achieved. This includes camera placement in order to reduce the camera-to-face angle, increase illumination and decrease the likelihood of having multiple people in the camera frame at once. These conditions however, could be difficult to replicate in real-life environments, making the prospect of FIAC in currently uncontrolled environments infeasible. In reality though, it is too difficult to assess the feasibility of FIAC in an operational environment based on the use of one algorithm.

Further research is required to be able to give more information about the performance level of the current FR technology in different conditions, possibly specific to a particular client/organisation interested in implementing the technology. The remaining footage obtained during the 2012 operational trial should be examined to assess all other areas and gain a better understanding of the true performance of FR in a range of different locations. This footage should be assessed with a range of different algorithms to gain a wider gauge of FR matching potential and give an indication of generic FIAC feasibility. The key locations should be examined more closely in follow-up trials where multiple cameras and supplementary lighting are used to provide potential environments for installation to determine the ceiling performance. Another area for future research is determining the feasibility of FIAC in diverse operational locations to assess the impact that different environments have on FR performance. This would contribute to a greater understanding of what aspects of an environment affects performance, for better and for worse.

Another aspect of FIAC that has been introduced into some FR algorithms since completing the 2012 Operational Trial is 'Tracking'. It would be beneficial to determine whether this additional tool affected FR performance. There are many aspects of this tool that would need assessment, which is outside the scope of this current research; however it is hypothesised that such a tool would be largely beneficial for many operational applications.

It is also important to look at how FR fits into a FIAC application and understand the role of the human operator in detecting and locating POIs. As previous research by DSTO suggests, a human operator is crucial to all 1:N FR systems and therefore it is also important to understand how the human performs in this application. This is a central point to the follow-through of a FR result. As the current research only focused on the algorithmic capability of FR functioning in a FIAC application, if the algorithms were found to be deployable, another study would need to be conducted to evaluate the processes that occur after a match is detected and further downstream affects.

In conclusion, although facial recognition algorithms for use in FIAC applications have shown improvement over the past decade, the feasibility of its deployment into unconstrained environments remains unclear. If the operational environment can be modified to become better suited for optimal image acquisition, then FIAC may be viable. It does however still depend on many variables that are specific to each environment. It is suggested that further research is required to process the data from the 2012 trial with updated and supplementary algorithms, as well as potentially run an additional trial in an alternate operational environment to gain a greater understanding of the feasibility of FIAC in the current day.

7. References

1. Introna, L.D. and H. Nissenbaum, *Facial recognition technology: A survey of policy and implementation issues.* Center for Catastrophe Preparedness and Response, New York University, 2009.
2. Bigdeli, A., B.C. Lovell, and S. Mau. *You, yes you: Welcome to the world of advanced surveillance*. 2011; Available from: http://theconversation.com/you-yes-you-welcome-to-the-world-of-advanced-surveillance-830.
3. Hampapur, A., Brown, L., Connell, J., Pankanti, S., Senior, A., Tian, Y. *Smart surveillance: applications, technologies and implications.* Information, Communications and Signal Processing, 2003. **2**(December): p. 1133-1138.
4. Bigdeli, A., B.C. Lovell, and S. Mau. *Face-in-the-crowd Biometrics: Here's looking secretly at you*. 2011; Available from: http://theconversation.com/face-in-the-crowd-biometrics-heres-looking-secretly-at-you-831.
5. Hancock, P.J., V. Bruce, and A.M. Burton, *Recognition of Unfamiliar Faces.* Trends in Cognitive Sciences, 2000. **4**(9): p. 330-337.
6. Shaokang, C., Mau, S., Harandi, M.T., Sanderson, C., Bigdeli, A., Lovell, B.C. *Face recognition from still images to video sequences: A local-feature-based framework.* EURASIP Journal on Image and Video Processing, 2011. **2011**.
7. Adler, A. and M.E. Schuckers, *Comparing Human and Automatic Face Recognition Performance.* Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2007. **37**(5): p. 1248-1255.
8. Phillips, P.J., Scruggs, W.T., O'Toole, A.J., Flynn, P.J., Bowyer, K.W., Schott, C.L., Sharpe, M. *FRVT 2006 and ICE 2006 large-scale experimental results.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010. **32**(5): p. 831-846.
9. Woodward Jr, J.D., Horn, C., Gatune, J., Thomas, A. *Biometrics: A look at facial recognition*. 2003, Santa Monica: RAND Corporation.
10. National Science and Technology Council, *Biometrics in Givernment Post 9-11; Advancing Science, Enhancing Operations*. 2008: Washington DC.
11. United States Congress Senate Committee on the Judiciary, *Biometric identifiers and the modern face of terror: new technologies in the global war on terrorism*, T. Subcommittee on

Technology, Editor. 2001, In Committee on th Judiciary of Technology, Terrorism and Government Information.

12. Yun, Y.W., *The '123' of Biometric Technology*, in *Laboratories for Information Technology Co-Chair, Biometrics Working Group of Security & Privacy Standards Technical Committee*. 2003.
13. Daugman, J., *Iris recognition at airports and border-crossings*, in *Encyclopedia of Biometrics*. 2009, Springer. p. 819-825.
14. McLindin, B., *Improving the Performance of Two-DImensional Facial Recognition Systems: The Development of a Generic Model for Biometric Technology Variables in Operational Environments*. 2005, University of South Australia.
15. Beveridge, J.R., Givens, G.H, Phillipa, P.J., Draper, B.A. *Factors that influence algorithm performance in the face recognition grand challenge.* Computer Vision and Image Understanding, 2009. **113**(6): p. 750-762.
16. Gross, R., J. Shi, and J. Cohn, *Quo vadis face recognition? The current state of the art in face recognition*. 2001, Technical Report TR-01-17, Carnegie Mellon University.
17. Podilchuk, C., Hulbert, W., Flachsbart, R., Barinov, L. *Face recognition for uncontrolled environments*. in *SPIE Defense, Security, and Sensing*. 2010. International Society for Optics and Photonics.
18. Crawford, M. *Facial Recognition Progress Report*. 2011; Available from: http://spie.org/x57306.xml.
19. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J. *The FERET evaluation methodology for face-recognition algorithms.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000. **22**(10): p. 1090-1104.
20. Chant, T.D. *The Limits of Facial Recognition*. NovaNext 2013 [cited 2014 04/02/2014]; Available from: http://www.pbs.org/wgbh/nova/next/tech/the-limits-of-facial-recognitionhttp://www.pbs.org/wgbh/nova/next/tech/the-limits-of-facial-recognition.
21. Klontz, J.C. and A.K. Jain, *A Case Study on Unconstrained Facial Recognition Using the Boston Marathon Bombings Suspects.* Michigan State University, Tech. Rep, 2013.
22. Paul S Gaertner, P., Curtis, N., Bowley, D., McLindin, B., Davy, J., Jaeger, J., & Bourn, S., (2003), *Identifying a Face in a Crowd: A Report on the Trial of One-to-Many Facial Recognition Systems at Sydney's Kingsford Smith Airport*, DSTO-DSTO-CR-0229, Defence Science and Technology Organisation, Edinburgh.

23. Butavicius, M., Graves, I., Hemming, I., Ivancevic, V., Johnson, R., Kaine, A., McLindin, B., Meaney, K., & Sunde., J, (2003), *CLASSIFIED TITLE*, DSTO-DSTO-CR-0292, Defence Science and Technology Organisation, Edinburgh.

24. Wolfhope, P., *Standoff Technology Integration and Demonstration Program: Biometric Optical Surveillance System Tests*, U.S. Department of Homeland Security, Editor. 2012, Science and Technology Directorate.

25. Savage, C., *Facial Scanning is Making Gains in Surveillance*, in *New York Times*. 2013: New York. p. A1.

26. Groebner, V., *Who Are You? Identification, Deception, and Surveillance in Early Modern Europe* in New York: Zone Books 2007.

27. Hindle, S., *Who, Me?.* History Workshop Journal (2009) **67**(1): 209-213

28. Blackburn, T., Butavicius, M., Graves, I., Hemming, D., Ivancevic, V., Johnson, R., Kaine, A., McLindin, B., Meaney, K., Smith, B., Sunde, J. *Biometrics technology review 2002*. 2003, Defence Science and Technology Organisation.

29. Gohringer, C., *The application of face recognition in airports.* Biometric Technology Today, 2012. **2012**(7): p. 5-9.

30. Spaun, N.A., *Facial comparisons by subject matter experts: Their role in biometrics and their training*, in *Advances in Biometrics*. 2009, Springer. p. 161-168.

31. Burton, A.M., D. White, and A. McNeill, *The Glasgow face matching test.* Behavior Research Methods, 2010. **42**(1): p. 286-291.

32. Megreya, A.M. and M. Bindemann, *Revisiting the processing of internal and external features of unfamiliar faces: The headscarf effect*. Perception, 2009. **38**(12): p. 1831-1848.

33. Behrman, B.W. and S.L. Davey, *Eyewitness identification in actual criminal cases: an archival analysis.* Law and human behavior, 2001. **25**(5): p. 475.

34. Kemp, R., N. Towell, and G. Pike, *When seeing should not be believing: Photographs, credit cards and fraud.* Applied Cognitive Psychology, 1997. **11**(3): p. 211-222.

35. Megreya, A.M. and A.M. Burton, *Matching faces to photographs: poor performance in eyewitness memory.* Journal of Experimental Psychology: Applied, 2008. **14**(4): p. 364.

36. Jenkins, R. and A. Burton, *Limitations in facial identification.* Justice of the Peace, 2008. **172**(40210): p. 4-6.

37. Bindemann, M., M. Avetisyan, and K.-A. Blackwell, *Finding needles in haystacks: Identity mismatch frequency and facial identity verification.* Journal of Experimental Psychology: Applied, 2010. **16**(4): p. 378.

38. Cole, S.A., *Suspect identities: A history of fingerprinting and criminal identification*. 2009, New York: Harvard University Press.
39. Kaluszynski, M., *The International Congresses of Criminal Anthropology: Shaping the French and International Criminological Movement, 1886-1914*, in *In: Becker, P and Wetzell,R.F (eds.) Criminals and their Scientists.* 2006, Cambridge University Press: Cambridge. p. 301-316.
40. Roark, D., A.J. O'Toole, and H. Abdi. *Human recognition of familiar and unfamiliar people in naturalistic video*. in *Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003. IEEE International Workshop on*. 2003. IEEE.
41. Graves, I., Butavicius, M., MacLeod, V., Heyer, R., Parsons, K., Kuester, N., McCormac, A., Jacques, P., Johnson, R. *The role of the human operator in image-based airport security technologies*, in *Innovations in Defence Support Systems-2*. 2011, Springer. p. 147-181.
42. Vast, R. and M. Butavicius, *A Literature Review of Face Recognition for Access Control: Human versus MAchine Solutions*, D.o. Defence, Editor. 2005, Defence Science and Technolohy Organisation: South Australi.
43. Pike, G., R. Kemp, and N. Brace. *The psychology of human face recognition*. in *Visual Biometrics (Ref. No. 2000/018), IEE Colloquium on*. 2000. IET.
44. Sinha, P., Balas, B., Ostrovsky, Y., Russell, R. *Face recognition by humans: Nineteen results all computer vision researchers should know about.* Proceedings of the IEEE, 2006. **94**(11): p. 1948-1962.
45. O'Toole, A., P. Phillips, and A. Narvekar. *Humans versus algorithms: Comparisons from the FRVT 2006*. in *Proceedings of the Eighth International Conference on Automatic Face and Gesture Recognition*. 2008.
46. Myers, J., *Surveillance Cameras*, in *Encyclopedia of Privacy*. 2007, Greenwood Press: Westport.
47. Barrett, J. *Where's our Bailout?* News 2001 [cited 2013 01/01/2013]; Available from: http://www.newsweek.com/wheres-our-bailout-149549+&cd=1&hl=en&ct=clnk&gl=au.
48. Bigdeli, A., Lovell, B.C., Sanderson, C., Shan, T., Chen, S. *Vision Processing in Intelligent CCTV for Mass Transport Security*. in *IEEE Workshop on Signal Processing Applications for Public Security and Forensics, 2007*. 2007. IET.
49. Burton, A.M., Miller, P., Bruce, P., Hancock, P.J.B., Henderson, Z. *Human and automatic face recognition: a comparison across image formats.* Vision Research, 2001. **41**(24): p. 3185-3195.
50. Chellappa, R., P. Sinha, and P.J. Phillips, *Face recognition by computers and humans.* Computer, 2010. **43**(2): p. 46-55.

51. Hill, H. and V. Bruce. *The effects of lighting on the perception of facial surfaces.* Journal of Experimental Psychology: Human Perception and Performance, 1996. **22**(4): p. 986.

52. Feingold, G.A. *Influence of environment on identification of persons and things.* J. Am. Inst. Crim. L. & Criminology, 1914. **5**: p. 39-51.

53. Furl, N., P.J. Phillips, and A.J. O'Toole, *Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis.* Cognitive Science, 2002. **26**(6): p. 797-815.

54. Tanaka, J.W., M. Kiefer, and C.M. Bukach, *A holistic account of the own-race effect in face recognition: Evidence from a cross-cultural study.* Cognition, 2004. **93**(1): p. B1-B9.

55. Meissner, C.A. and J.C. Brigham, *Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review.* Psychology, Public Policy, and Law, 2001. **7**(1): p. 3.

56. Walker, P.M. and M. Hewstone, *A developmental investigation of other-race contact and the own-race face effect.* British Journal of Developmental Psychology, 2006. **24**(3): p. 451-463.

57. Rohan, M. *New Report Next Generation Biometric Technologies Market (2012-2017) by MarketsandMarkets*. 2012; Available from: http://www.prweb.com//releases/next-generation-biometric/technologies-market/prweb9666400.htm.

58. Xiao, Q., *Technology review-biometrics-technology, application, challenge, and computational intelligence solutions.* Computational Intelligence Magazine, IEEE, 2007. **2**(2): p. 5-25.

59. Heyer, R., *Biometrics Technology Review 2008*. 2008: Defence Science and Technology Organisation.

60. Jain, A. and S. Pankanti, *Automated fingerprint identification and imaging systems.* Advances in Fingerprint Technology, 2001. **2**.

61. Komarinski, P., *Automated fingerprint identification systems (AFIS)*. 2005, London: Academic Press.

62. National Science and Technology Council, *Biometrics Testing and Statistics*, Homeland and National Security, Editor. 2006, NSTC: Washington.

63. Nanavati, R., *Biometric Border Security Evaluation Framework (Biometrique Cadre D'evaluation de la Securite des Frontieres)*. 2011, Defence Research and Development Canada: Ottawa.

64. Daugman, J., *How iris recognition works.* Circuits and Systems for Video Technology, IEEE Transactions on, 2004. **14**(1): p. 21-30.

65. Matey, J.R., Naroditsky, O., Hanna, K., Kolczynski, R., Lolacono, D.J., Mangru, S., Tinker, M., Zappia, T.M., Zhao, W.Y., *Iris on the move: Acquisition of images for iris recognition in less constrained environments.* Proceedings of the IEEE, 2006. **94**(11): p. 1936-1947.

66. National Science and Technology Council, *Face Recognition*, Department of Homeland Security, Editor. 2006, NSTC: Washington.
67. O'Toole, A.J., Phillips, P.J., Jiang, F., Ayyad, J., Penard, N., Abdi, H., *Face recognition algorithms surpass humans matching faces over changes in illumination.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2007. **29**(9): p. 1642-1646.
68. Spaun, N.A., *Face Recognition in Forensic Science*, in *Handbook of Face Recognition*. 2011, Springer. p. 655-670.
69. Gates, K., *The past perfect promise of facial recognition technology*. 2004.
70. Chellappa, R., C.L. Wilson, and S. Sirohey, *Human and machine recognition of faces: A survey.* Proceedings of the IEEE, 1995. **83**(5): p. 705-741.
71. Phillips, P.J., Flynn, P.J, Scruggs, T., Bowyer., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W. *Overview of the face recognition grand challenge*. in *IEEE computer society conference on Computer vision and pattern recognition, 2005*. 2005. IEEE.
72. Grother, P. and M. Ngan, *Performance of Face Identification Algorithms*. 2014, National Institute of Standards and Technology
73. Tan, X., Chen, S., Zhou, Z-H., Zhang, F., *Face recognition from a single image per person: A survey.* Pattern Recognition, 2006. **39**(9): p. 1725-1745.
74. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A., *Face recognition: A literature survey.* Acm Computing Surveys (CSUR), 2003. **35**(4): p. 399-458.
75. Huang, T., Z. Xiong, and Z. Zhang, *Face recognition applications*, in *Handbook of Face Recognition*. 2011, Springer: London. p. 617-638.
76. Stikeman, A., *Recognizing the enemy.* Technology Review, 2001. **104**(10): p. 48-49.
77. Australian National Audit Office, *Processing and Risk Assessing Incoming International Air Passengers*. 2012, Australian Customs and Border Protection Service: Australian Capital Territory.
78. Grother, P.J., G.W. Quinn, and P.J. Phillips, *Report on the evaluation of 2D still-image face recognition algorithms.* NIST interagency report, 2010. **7709**: p. 106.
79. Jenkins, R. and D. White. *Commercial face recognition doesn't work*. in *Symposium on Bio-inspired Learning and Intelligent Systems for Security, 2009*. 2009. IEEE.
80. Valencia, V.S. *Biometric testing: It's not as easy as you think*. in *Biometric Consortium Conference*. 2003. Arlington VA.
81. Bonner, J., *Looking for Faces in the Super Bowl Crowd.* Access Control & Security Systems Magazine, 2001.

82. Greene, L. *Face scans match few suspects*. St Petersburg Times Online 2001; Available from: http://www.sptimes.com/News/021601/TampaBay/Face_scans_match_few_.shtml.

83. Greene, T.C. *Face Recognition useless for crowd surveillance*. The Register 2001; Available from: http://www.theregister.co.uk/2001/09/27/face_recognition_useless_for_crowd/.

84. Dennis, B. *Ybor cameras won't seek what they never found*. 2003 [cited 2012 25/08/2013]; Available from: http://www.sptimes.com/2003/08/20/Hillsborough/Ybor_cameras_won_t_se.shtml.

85. Titsworth, T., *More than face value: airports and multimedia security.* MultiMedia, IEEE, 2002. **9**(2): p. 11-13.

86. Brooks, M., *Face Off*, in *New Scientist*. 2002. p. 28-42.

87. Soldatov, A. and I. Borogan, *A face in the crowd: the FSB is watching you*, in *OpenDemocracy*. 2011, openDemocracy Limited: London.

88. Pretzel, K. and T. Lotz, *Face Recognition as a search tool "Foto Fahnsung" Final Report*. 2007, German Federal Criminal Police Office (BKA): Mainz, Germany.

89. Rawlinson, T., A. Bhalerao, and L. Wang, *Principles and methods for face recognition and face modelling*, in *In: Handbook of Research on Computational Forensics, Digital Crime and Investigation: Methods and Solutions*, L. Chang-Tsun, Editor. 2010, IGI Global: Warwick. p. 55-78.

90. Givens, G., Beveridge, J.R., Draper, B.A., Grother, P., Phillips, P.J., *How features of the human face affect recognition: a statistical comparison of three face recognition algorithms*. in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004*. 2004. IEEE.

91. Jafri, R. and H.R. Arabnia, *A Survey of Face Recognition Techniques.* Journal of Information Processing Systems, 2009. **5**(2): p. 41-68.

92. Adini, Y., Y. Moses, and S. Ullman, *Face recognition: The problem of compensating for changes in illumination direction.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1997. **19**(7): p. 721-732.

93. Mansfield, A.J. and J.L. Wayman, *Best practices in testing and reporting performance of biometric devices*. 2002: Centre for Mathematics and Scientific Computing, National Physical Laboratory Teddington, Middlesex, UK.

94. Organisation, I.S., *Information Technology- Biometric Performance Testing and Reporting*. 2006, ISO/IEC: Switzerland.

95. D'agostino, R.B., A. Belanger, and R.B. D'Agostino Jr, *A suggestion for using powerful and informative tests of normality.* The American Statistician, 1990. **44**(4): p. 316-321.

96. Pallant, J., *SPSS survival manual: A step by step guide to data analysis using SPSS*. 2010, England: Open University Press.
97. Cohen, J., *Statistical power analysis for the behavioural sciences (2nd Edition) (1988). New Jersey: Lawrence Erlbaum.*
98. Morgan, S.E., T. Reichert, and T.R. Harrison, *From numbers to words: Reporting statistical results for the social sciences*. 2002, Boston: Allyn and Bacon.
99. Gaertner, P., *Title Restricted*, Department of Defence, Editor. 2003, Defence Science and Technology Organisation: South Australia.
100. McLindin, B., *Biometrics Research Program Imaging Trial*. 2006, Defence Science and Technology Organisation: South Australia.